



Search for new physics signatures in ATLAS data using autoencoders

Master-Thesis

zur Erlangung des akademischen Grades

Master of Science

vorgelegt von

FREDERIK STROTHMANN

Wuppertal, den 03. März 2020



Erstgutachter: Prof. Dr. Wolfgang Wagner

Zweitgutachter: Dr. Dominic Hirschbühl

Abstract

Searches for new physics beyond the Standard Model are mostly driven by theories predicting new particles or processes. Due to the theoretical bias of these searches, the investigated phase space is very narrow. Thus, a new signal might not be found even if it occupies a phase space close to the investigated one. Autoencoders offer a semi supervised and bias free method to search for new signals in any region. They are a special type of neural networks, that reproduces the input using a lower dimensional representation. Autoencoders are trained in a semi supervised manner, which allows them to detect anomalies while only being trained on known processes of the Standard Model. This master thesis evaluates autoencoders as a method to search for new physics signatures, using data from pp collisions at $\sqrt{s}=13$ TeV measured by the ATLAS detector. To evaluate autoencoders, a region that is dominated by $t\bar{t}$ events, with a minor fraction of single top tW events is used. For that region a discriminant called mini-max $m(lb)$ already exists. A detailed comparison between the two algorithms is carried out. The autoencoder learns a partial representation of mini-max $m(lb)$ and additionally selects a region with increased discrepancy, that is not included by mini-max $m(lb)$. Furthermore, an autoencoder is applied to a region with a discrepancy between data and simulation, where the origin of the discrepancy is unknown. The autoencoder suggests that the discrepancy originates from a shape mismatch and/or a scaling error.

Contents

1	Introduction	1
2	Experimental and theoretical context	3
2.1	The Standard Model of particle physics	3
2.1.1	Elementary particles	3
2.1.2	Interactions	4
2.1.3	Open questions of the Standard Model	5
2.2	Proton-Proton Collisions	5
2.3	Top quark production at the LHC	6
2.4	The ATLAS detector at the LHC	8
2.4.1	Large Hadron Collider	8
2.4.2	The ATLAS detector	10
2.5	Object reconstruction	13
2.6	Kinematic reconstruction	14
3	Statistical methods	17
3.1	Neural Networks	17
3.2	Neural Networks for classification	19
3.3	Autoencoder	20
4	Searches beyond the standard model	23
4.1	Search for new physics with a prior model	23
4.2	Search for new physics without a prior model	26
4.3	Using Autoencoders for new physics searches	26
5	Analysis of Autoencoders under controlled conditions	29
5.1	Motivation	29
5.2	Data and simulated samples	29
5.3	Object definition and event selection	31
5.3.1	Object definition	31
5.3.2	Event Selection	32
5.4	Kinematic properties	32
5.5	Analysis strategy	36
5.6	mini-max $m(\text{bl})$	37
5.6.1	Variables after mini-max $m(\text{bl})$ cut	39
5.7	Autoencoder	41
5.7.1	Variables and preprocessing	41

5.7.2	Architecture	41
5.7.3	Results	43
5.8	Classifier	46
5.8.1	Preprocessing	47
5.8.2	Architecture	47
5.8.3	Training	47
5.8.4	Results	47
5.8.5	Conclusion	50
5.9	Comparison	51
5.10	Conclusion	58
6	Application of Autoencoders for $t\bar{t}$+jets	61
6.1	Data and simulated samples	61
6.1.1	Monte Carlo samples	61
6.1.2	Object definition	62
6.2	Event selection	63
6.3	Kinematic properties	64
6.4	Autoencoder	66
6.5	Result and conclusion	68
7	Conclusion and outlook	69
	Bibliography	71
A	Appendix	77
A.1	Kinematic properties (Section 5.4)	77
B	Appendix	82
B.1	Kinematic properties (Section 6.3)	82

1 Introduction

The Standard Model of particle physics (SM) is the best description of elementary particles and their interactions to date. But there are still phenomena, that are not explained by the SM. These are phenomena like dark matter or B-anomalies observed by LHCb. Theories that try to explain these phenomena introduce new particles and interactions. Searches for these new particles and interactions are performed at the Large Hadron Collider (LHC).

Theories for new particles and processes impose restrictions on the phase space that can be used by an analysis. Due to these limitations, any analysis based on these assumptions is narrow in scope and might not be able to find the signal even if it is in a phase space close to the observed one, due to its theoretical bias.

With the rise of Deep Learning (DL) in the past few years, Neural Networks (NNs) have been incorporated into many analyses. In most cases NNs are used for classification or regression in a supervised manner, which introduces a bias into the analyses. Autoencoders (AEs) offer a semi supervised option, that is free from any theoretical bias to search for new signals in any region. Semi supervised means that the AE only needs the simulated events of the processes which are predicted by the SM. The goal of this thesis is to investigate the usability of AEs for the search of new physics signal using kinematic variables.

The structure of this work is as follows: Chapter 2 gives an overview, on the scientific context of this thesis, characterized by the SM, the LHC, the ATLAS detector and the field of top quark physics in particular. Chapter 3 introduces the fundamentals of NNs and gives a description of NN classifiers and AEs. The fourth chapter gives a brief overview for searches beyond the SM, that are either based on a model of the new physics or not. The main part of this work is covered in Chapter 5. At first, the used data and simulated samples are introduced. Afterwards the physics-motivated variable mini-max $m(lb)$ is introduced, which is used as a benchmark for the AE result. Furthermore, the AE, its training, and the results are described. To have a comparison between a semi supervised and fully supervised algorithm, a NN classifier is trained for comparison. Then a detailed analysis between the differences of mini-max $m(lb)$ and the AE is done.

In Chapter 6 an AE is used to investigate an observed discrepancy between simulation and measured data. The goal is to investigate if the discrepancy between simulation and measured data originates from a new signal or a false scaling and/or shape miss match. Therefore, at first an introduction is given, then data and simulated samples are introduced. Afterwards a comparison between measured data and simulation is done for the kinematic variables. Next, the AE and its training is described. At the end the results are presented

and a conclusion is drawn. In the last chapter a conclusion on the whole work and an outlook for future possibilities for the usage of AEs in particle physics is given.

2 Experimental and theoretical context

In this chapter an overview of the SM and the ATLAS detector at the LHC is given. At the heart of understanding and testing the SM are proton-proton collisions which are done at particle colliders like the LHC.

2.1 The Standard Model of particle physics

The SM describes the elementary particles and their interactions [1, 2]. The model includes three fundamental forces: The electromagnetic, weak, and strong force (gravitation is not included). It was developed in the 1960s and 1970s and has last been confirmed in 2012 with the discovery of the Higgs boson by the ATLAS and CMS experiments [3, 4]. Even though the SM had great success there are phenomena that are not yet explained. The precise measurement of the SM predictions and the search for new physics beyond the SM are tasks of experiments probing the SM, like ATLAS.

2.1.1 Elementary particles

Elementary particles are divided into two groups, fermions and bosons, according to their spin. Fermions have a half-integer spin and follow the Fermi-Dirac statistic. They are subject to the Pauli exclusion principle, which states that no two fermions can occupy the same quantum state. Fermions are further differentiated into leptons and quarks, where leptons have a weak isospin, but no color charge like quarks do. Leptons exist in two main classes, the electron like leptons which have an electrical charge and the neutrinos which have no electrical charge. There are six quarks (up, down, charm, strange, top, bottom) and six leptons (electron, muon, tau, electron neutrino, muon neutrino and tau neutrino). Higher generations of fermions have the same properties as the lower generations except for a higher mass. Only the lightest fermions are stable and thus the heavier fermions decay into the lighter ones. All fermions have an anti-particle that has the same properties except for an opposite charge.

Bosons have an integer spin and obey the Bose-Einstein statistic. Gauge bosons are the mediators of the fundamental forces. The $W^\pm$ bosons have a negative or positive charge, while photons, gluons, Z, and Higgs bosons are electrically neutral. The Higgs boson is the only scalar (spin 0) particle and plays an important role in explaining the particle masses. In Figure 1 a graphical overview of the SM particles is given.

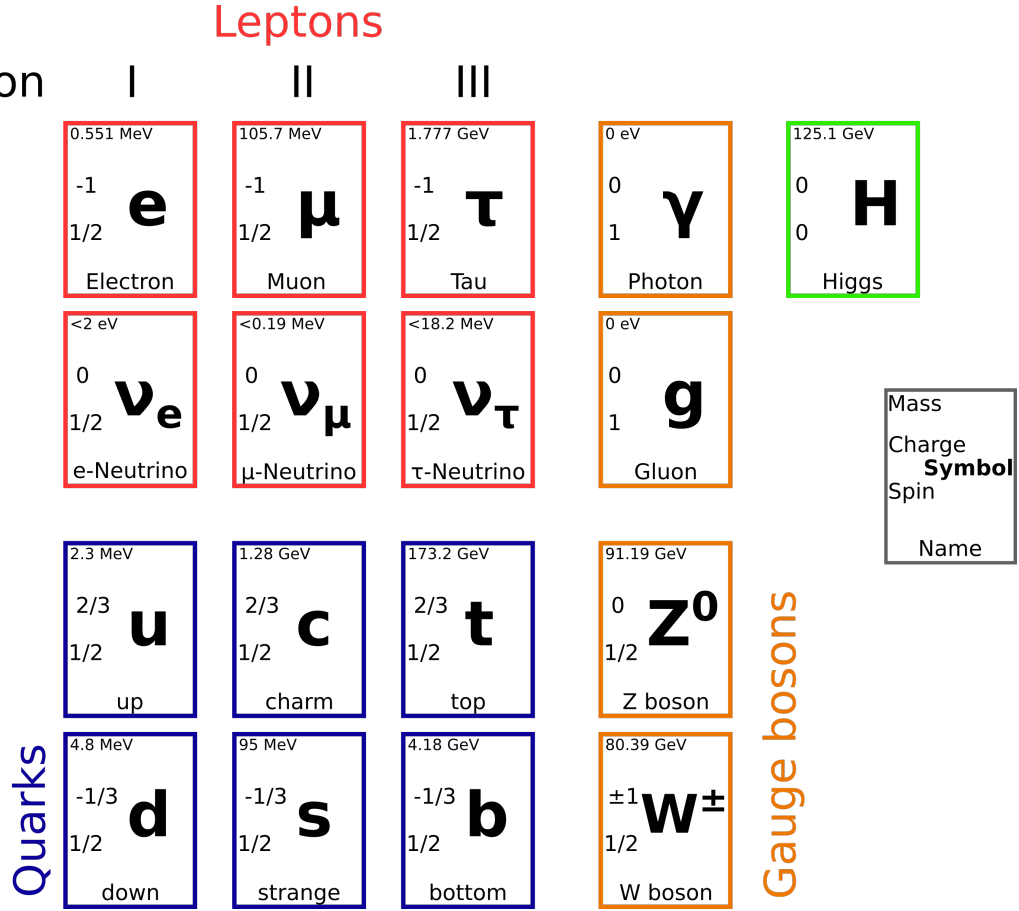


Figure 1: Overview of the SM particles. Properties for the particles are taken from Ref. [5]

2.1.2 Interactions

In the SM, the three forces (electromagnetic, weak, and strong) are mediated through gauge bosons. The electromagnetic force, described by Quantum electrodynamics (QED), is mediated by massless photons. The range of the electromagnetic force is infinite due to no self couplings of photons. The weak force is mediated by the Z boson (neutral current) and the $W^\pm$ bosons (charged current). They couple to the weak isospin of a particle and thus affect all elementary particles. The weak force maximally violates parity, that means charged currents only couple to left-handed fermions. Charged currents can change the quark flavor as well as the electrical charge. The transition probabilities for quark flavor changes under charged currents is described by the Cabbibo-Kobayashi-Maskawa-matrix (CKM-matrix).

$$V_{\text{CKM}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \quad |V_{ij}| = \begin{pmatrix} 0.974 & 0.225 & 0.004 \\ 0.225 & 0.973 & 0.041 \\ 0.009 & 0.041 & 0.999 \end{pmatrix} \quad (1)$$

The CKM-matrix is a unitary matrix, which means flavor changing neutral currents (FCNC) are excluded at tree-level in the SM. The electromagnetic and the weak force are unified in the electroweak force.

The theory describing the strong force is called quantum chromodynamics (QCD). The strong force is mediated by massless gluons that carry a color charge. It acts on the color charge of a particle and thus couples to all quarks and other gluons. The strong force increases with distance between the quarks. Due to the confinement it is not possible to free quarks. When quarks get separated at a distance of ≈ 1 fm, it is energetically more efficient to create a new quark-antiquark pair. This process is called hadronization. Quarks only exist in color neutral states as hadrons. Hadrons can consist of two quarks (mesons) or three quarks (baryons).

2.1.3 Open questions of the Standard Model

The SM has proven to be very successful and describes most of the known phenomena, but there are phenomena not described by the SM. Examples of such phenomena are gravity, neutrino oscillations, and dark matter.

Gravitation is the weakest of the four fundamental forces, it is multiple magnitudes smaller as the forces described in the SM. At the usual energy scale of the processes of the SM the gravitational effect can be neglected. But at energies near the Planck scale gravitation becomes important.

The SM assumes that the mass of neutrinos is zero, which would not allow them to oscillate between the neutrino flavors. Experiments such as Super-Kamiokande have shown that neutrinos oscillate [6] which means that their mass has to be above zero.

Cosmological observations show that the biggest part of the visible universe is made up by dark matter. The SM does not provide a particle that would explain its origin.

Many theories beyond the Standard Model (BSM) try to explain these phenomena by introducing new particles or processes.

2.2 Proton-Proton Collisions

In order to probe the SM and its predictions, measurement data from colliders like LHC are compared against simulation from Monte Carlo (MC) generators [7]. The proton-proton collisions at the LHC happen in three steps: 1. hard scattering from two protons that collide (short distance, high energy), 2. radiation and parton showers (long distance, low energy), 3. hadronization. These three steps can be looked at independently due to the factorization theorem [8], which states that processes at long and short distance can be separated at a factorization scale μ_F .

A proton consists of quarks and gluons, called partons. In proton-proton collisions the partons interact with each other. The parton distribution function (PDF) $f_i(x, \mu^2)$ describes the probability of the i -th parton carrying the fraction x of the momentum of the proton. The PDFs are obtained by fits to data from deep inelastic scattering (DIS) and hadron collisions [9]. The PDFs for different partons at energy scales of $\mu^2 = 10^4 \text{ GeV}^2$ and $\mu^2 = 10 \text{ GeV}^2$ are shown in Figure 2.

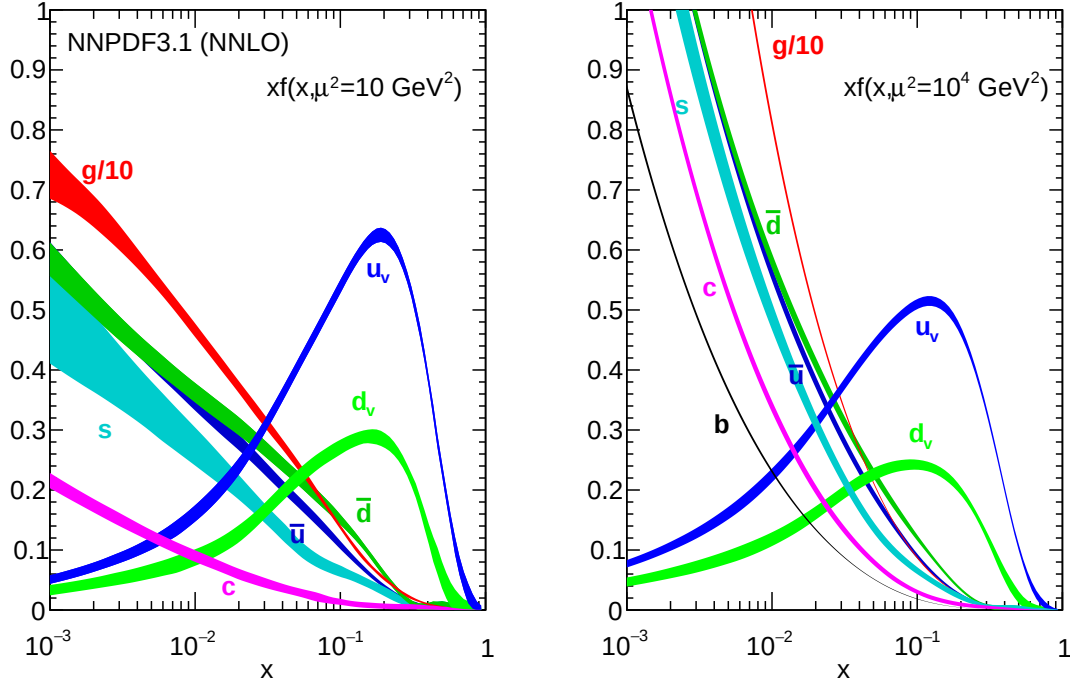


Figure 2: The NNPDF3.1 parton distribution function [10] for protons evaluated at $\mu^2 = 10 \text{ GeV}^2$ (left) and $\mu^2 = 10^4 \text{ GeV}^2$ (right).

The hard scattering is described by the interaction between partons. With the PDFs the cross section for hadron collisions can be calculated. For simulating the hard scattering process matrix element (ME) event generators are used. In parton showers, the partons emit gluons which themselves emit gluons and split into quark-antiquark pairs. When the energy of the partons is low enough, the partons build hadrons due to the effect of confinement, called hadronization.

2.3 Top quark production at the LHC

The top quark is the heaviest of the SM quarks with a mass of $173.0 \pm 0.4 \text{ GeV}$ [5], and has a mean lifetime of $3.29 \cdot 10^{-25} \text{ s}$. This is below the timescale for interactions of the strong force and therefore it cannot hadronize like all other quarks. The top quark was

discovered in 1995 by the Collider Detector at Fermilab (CDF) and $D\bar{D}$ experiment [11, 12]. Due the large mass of the top quark, only very powerful accelerators are capable of top quark production.

Top quark production can be divided into two categories, top-quark-pair and single-top-quark production. Pair production can occur either through gluon fusion or quark-antiquark annihilation, Figure 3 shows the corresponding leading order Feynman diagrams. Single top quark production happens by the weak interaction. The single top production can happen in three different modes called s-channel, t-channel and tW -channel. Figure 4 shows the production of a top quark in the tW channel.

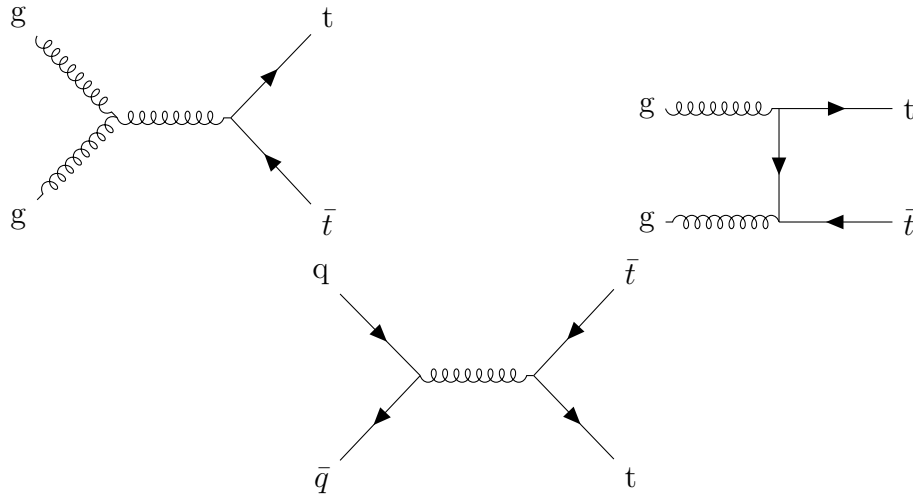


Figure 3: Leading order Feynman diagrams for $t\bar{t}$ production by gluon-gluon fusion (upper two) and quark-antiquark annihilation. At LHC for $\sqrt{s} = 14$ TeV gluon-gluon fusion makes up the majority of pair production with 90%, the other 10% come from quark-antiquark annihilation at the leading order (LO) [5].

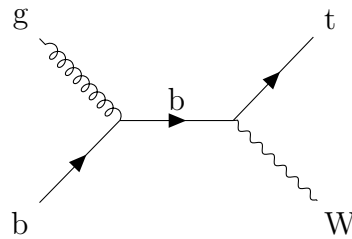


Figure 4: Leading order Feynman diagram for top-quark production in association with a W boson (tW channel).

Top quarks could also originate from decays of heavier particles predicted by BSM theories. One example of these particles is the vector-like top quark (VLT), that would decay into a top quark and a Z or Higgs boson [13].

The top quark decays into a Wb pair. The W boson decays further into a charged lepton and the corresponding neutrino or two quarks. Thus, final states of $t\bar{t}$ decays at LO can be split into three categories. The first one happens in 45.7% of all cases and is called all hadronic channel, the second appears in 43.8% of the decays and is called leptons+jets channel and the last one is called the dilepton channel, which makes up 10.5% of the decays [5]. Figure 5 shows the Feynman diagram of the possible decays.

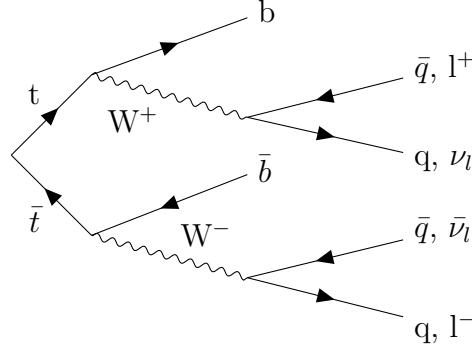


Figure 5: Feynman diagram for $t\bar{t}$ decay at LO. The three cases described above are achieved by selecting the combination of quarks and leptons accordingly.

Due to the increased amount of data produced during Run 2 at the LHC, more complex processes like $t\bar{t} + \gamma$ and $t\bar{t} + Z$ can be studied in more detail in the future.

2.4 The ATLAS detector at the LHC

Particle accelerators are the essential tool of particle physics and offer a way to look at the smallest building blocks of matter and their constituents. The LHC is the biggest accelerator build so far and ATLAS one of the experiments detecting the particle collisions at LHC. Both are now described in more detail.

2.4.1 Large Hadron Collider

The LHC [14] is a particle collider of the European Organization for nuclear research (CERN), located near Geneva, Switzerland. It has a circumference of 26.7 km and was built in a tunnel 100 m below the surface. The tunnel was originally constructed for the Large Electron-Positron Collider (LEP) accelerator from 1984-1989. The center of mass energy $\sqrt{s}$ of the collision can be up to 14 TeV, which makes it the most powerful particle accelerator in the world. Two runs have been done so far. The first one was from 2011-2012 with $\sqrt{s} = 7$ TeV. From 2013-2015 the accelerator was shutdown to prepare and upgrade it for the next run at $\sqrt{s} = 13$ TeV from 2015-2018.

The LHC has four interactions points, where detectors are placed (ALICE [15], ATLAS [16], CMS [17], LHCb [18]). ALICE is specialized for heavy ion collision measurements. ATLAS and CMS are multi-purpose detectors, that are used for SM measurements as well as BSM searches. LHCb is designed to study b-hadron physics.

The accelerator complex consists of many accelerators, where the smaller accelerators are used as pre-acceleration stages, before the particles are injected in the next accelerator. A schematic of the setup can be seen in Figure 6.

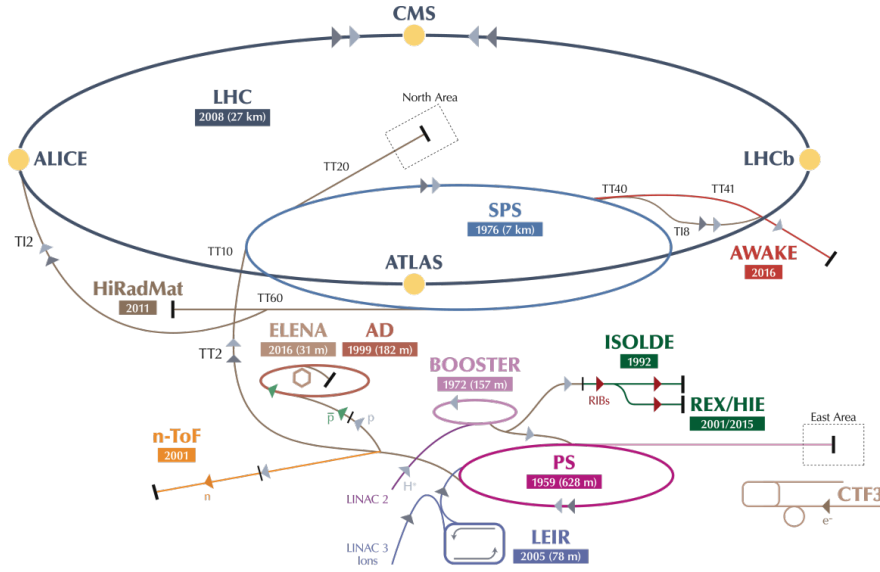


Figure 6: Schematic figure of the CERN accelerator complex [19].

At first protons are extracted from hydrogen atoms with an electric field. The extracted protons are then accelerated to an energy of 50 MeV with the linear accelerator LINAC2. Then the protons are accelerated to 1.4 GeV with the Proton Synchrotron Booster (PSB) and sent into the Proton Synchrotron (PS) that accelerates them to 25 GeV. Before the protons are injected into the LHC, the Super Proton Synchrotron (SPS) accelerates the protons to 450 GeV. In the LHC the protons are accelerated in opposite directions with an energy of 6.5 TeV.

In the LHC protons are bunched with a theoretical limit of 3564 bunches, where the bunches are 25 ns apart which gives a maximum collisions frequency of 40 MHz. Due to technical limitations, the maximum number of bunches is 2556 with each bunch containing $1.5 \cdot 10^{11}$ protons. The luminosity achieved by LHC is $\mathcal{L} = 2 \cdot 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, where $\mathcal{L}$ is defined as

$$\mathcal{L} = \frac{n_b N_{b,1} N_{b,2} f_r}{4\pi\sigma_x\sigma_y}, \quad (2)$$

where n_b is the number of bunches per beam, $N_{b,1/2}$ the number of particles per bunch, f_r the revolution frequency, and $\sigma_{x,y}$ the width of the beam profile. The event rate dN/dt is given by

$$\frac{dN}{dt} = \mathcal{L} \cdot \sigma, \quad (3)$$

where $\mathcal{L}$ is the instantaneous luminosity and σ the cross section. The integrated luminosity is defined as the integral of the luminosity over time. With the integrated luminosity the number of events is given by

$$N = \sigma \int dt \mathcal{L}. \quad (4)$$

Figure 7 shows the delivered luminosity as function of time for each year the LHC was running (left) and the total integrated luminosity of the LHC (right) in Run 2.

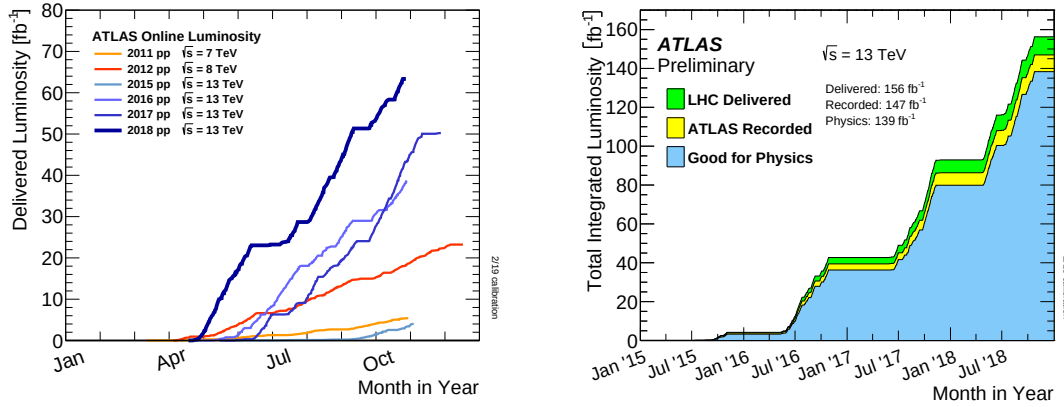


Figure 7: Delivered luminosity for pp collisions as a function of Month in a Year (left) and the total integrated luminosity for data taken from 2015-2018 (right) [20]

2.4.2 The ATLAS detector

ATLAS was originally an acronym for **A Toroidal LHC ApparatuS**, but is now only used as a proper name [16]. It was built for testing the known SM and searching for new physics beyond the SM. In 2012 ATLAS achieved its biggest success so far, with the discovery of

the Higgs boson [3, 4] in cooperation with the CMS collaboration.

Coordinate system

The coordinate system of ATLAS is centered at the interaction point in the center of the detector. The x-axis points towards the center of the LHC ring, the y-axis points upwards and the z-axis points along the beam pipe. In the transverse plane cylindrical coordinates are used, where ϕ is the azimuthal angle and r is the radius. The polar angle θ is measured with respect to the z-axis. In most cases the rapidity

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right) \quad (5)$$

is used. E is the energy of the particle and p_z is the z-component of the momentum. For particles that have a much higher energy than mass the rapidity can be approximated by the pseudorapidity

$$\eta = -\ln \left(\tan \left[\frac{\theta}{2} \right] \right). \quad (6)$$

The distance between two particles ΔR is defined as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}, \quad (7)$$

where $\Delta\eta$ and $\Delta\phi$ are the differences between the pseudorapidity and the azimuthal angle of the two particles.

Overview

ATLAS [16] is a general purpose detector. It has a cylindrical structure with a length of 44 m and a radius of 25 m. Due to the cylindrical structure, the detector nearly covers an 4π angle. The central part is called the barrel region and consists of multiple layers of sub-detectors, that are build cylindrical around the beam pipe. The parts before and behind the central part of the detector are called end-cap region and forward region. The components of the detector can be split into four parts: the inner detector, the calorimeter system, the muon system, and the trigger system. Figure 8 shows an overview of the ATLAS detector.

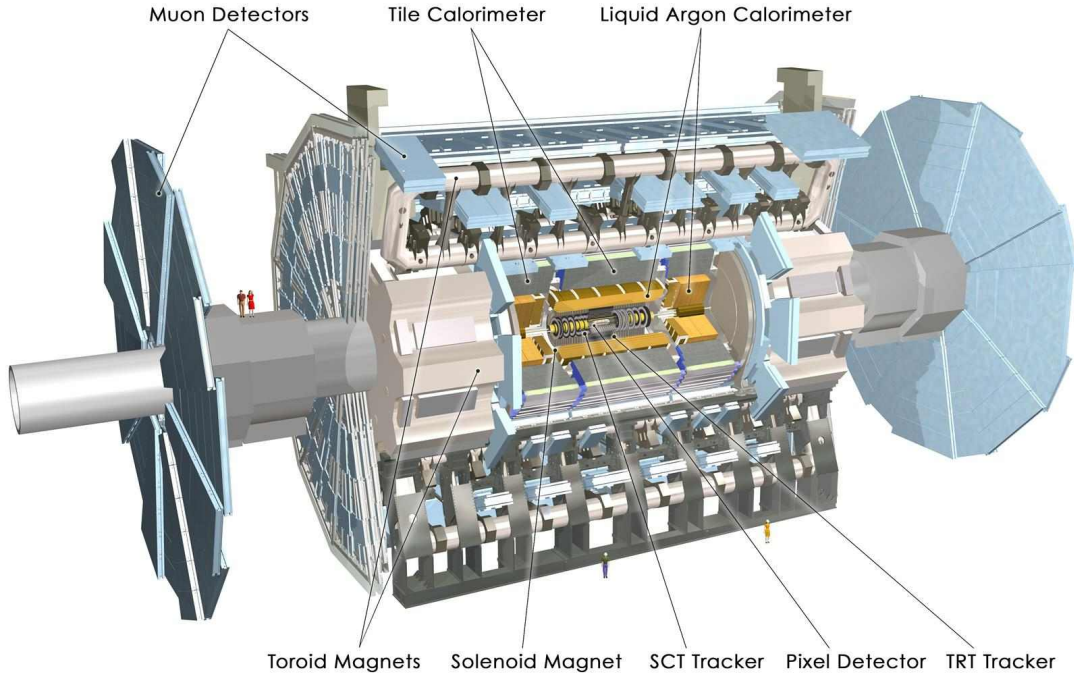


Figure 8: Overview of the ATLAS detector [21]

The Inner Detector

The inner detector (ID) is the part closest to the beam pipe and is surrounded by a superconducting solenoid. The solenoid provides a 2 T magnetic field, that is used to bend the path of the particles based on their charge. The ID measures the direction, momentum, and charge of the particles. It has three parts: the Pixel detector, the Semiconductor Tracker (SCT), and the Transition Radiation Tracker (TRT). The Pixel and SCT are used to reconstruct the origin of the particles and cover a range of $|\eta| < 2.5$. During the long shutdown 1 (February 2013–March 2015), the Insertable B-Layer was installed [21] to improve the performance of the tracking system. The TRT provides additional spatial measurements and information about the particle type and covers a region of $|\eta| < 2.0$.

The calorimeter system

The calorimeter system measures the energy of the particles. It surrounds the ID and covers a range of $|\eta| < 4.9$. The calorimeter consists of multiple parts that are used for measuring different particles. The electromagnetic calorimeter is used for precise measurements of electrons and photons. It consists of three parts: one surrounding the ID, one at the end-caps, and one in the forward regions. All three use liquid argon as active material. The absorber material for the center one surrounding the ID is lead, the outer ones use copper. The hadronic calorimeters at the end-cap and forward region use liquid argon as active material, the absorber at the end-cap is copper and in the forward region

tungsten. The central hadronic calorimeter, the Tile Hadronic Calorimeter (TileCal) uses scintillating tiles as active material and steel as absorber material.

The muon system

The muon system is designed to measure the momentum of muons even at high $p_t=1$ TeV, with a precision of 10%. To measure the track of the muons they are deflected by superconducting toroidal magnets, within a region of $|\eta|<2.7$. The measurement is done with four different types of muon chambers. Resistive Plate Chambers (PRCs) and Thin Gap Chambers (TGCs) are used for triggering. Monitored Drift Tubes (MDTs) and Cathode Strip Chambers (CSCs) are used for precise track measurement.

The trigger system

Due to technical limitations and the amount of data produced, not all events can be stored. To select events with interesting properties ATLAS uses a two stage trigger system [22]. The first stage, called L1 trigger, uses information from the calorimeters and the muon system. The information from these two sources is combined in the Central trigger of Level-1 to select events. This process takes $2.5 \mu\text{s}$, which gives a frequency of 100 kHz. The information of the L1 trigger are further propagated to the software based High Level Trigger (HLT) reducing the event rate to 1kHz. The triggered events are finally stored to disk for further analysis.

2.5 Object reconstruction

The collision products and their decay products leave a signature in the detector, from which the underlying partonic process can be reconstructed. Figure 9 shows in which part the different particles create a signal. Neutrinos are not directly visible to the detector, but due to conservation of momentum their transverse momentum can be reconstructed. For the colliding protons the transverse momentum is zero, so the sum over all transverse momenta of the collision products has to be zero as well. When summing over all the transverse momenta of all measured particles, the missing part to make the sum zero belongs to neutrinos or possibly a new particle invisible to the detector. This part is called missing transverse momentum E_T^{miss} . Electrically charged particles are deflected by the superconducting solenoid surrounding the ID. Photons and electrons are stopped and detected in the electromagnetic calorimeter. A distinction between the two is possible with a combination of the information from the ID and the calorimeter. Muons are not stopped but create a signal in the muon spectrometer. Hadrons deposit their energy in the hadronic calorimeter. The energy deposits in the calorimeter are used by a cluster-algorithm to cluster jets. A very important type of jet is the b-jet, because of the long

lifetime of b-hadrons containing a b-quark the distance between the decay point and the primary vertex can be measured. For identifying a b-jet a b-tagging algorithm is used. After the reconstruction an overlap removal algorithm is used to remove ambiguities, in the case a signal was identified as two different particles. The reconstructed objects are later used in analyses, for example to reconstruct top quarks or W bosons.

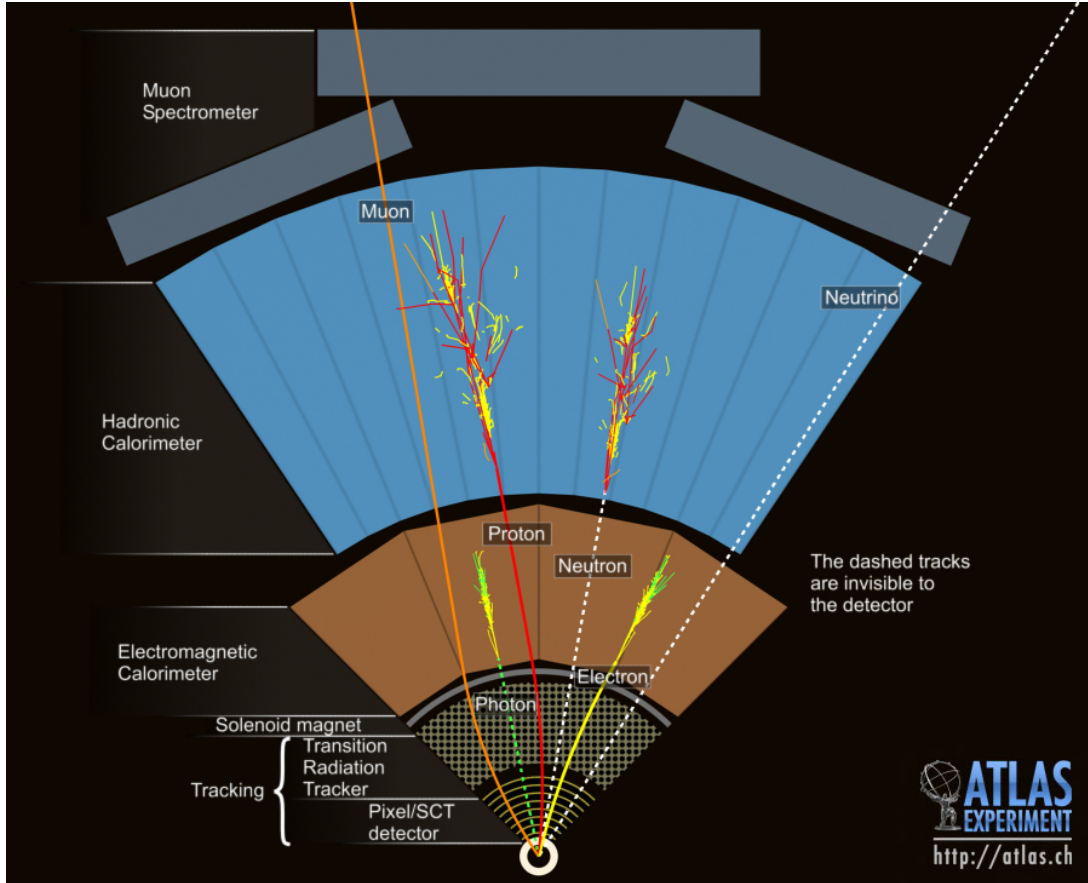


Figure 9: Overview of the object reconstruction in the different parts of the detector [23].

2.6 Kinematic reconstruction

Particles like the top quark cannot be measured directly, but their decay products can be. The top quark decays in 100% of the cases via $t \rightarrow Wb$, the W boson decays further. From the measured 4 vectors of the decay products the 4 vector of the particle that decayed into them can be calculated, this is called kinematic reconstruction. Different decays can have similar final states and only information from the final states can be measured. The process that should be measured is called signal, whereas processes that have similar final states, but should not be measured are called background. The separation of signal and background is not a trivial task and requires a good understanding of the physics of the underlying processes and multivariate methods. The next section will present the main

statistical method used in this work, namely AEs which present a new tool for separating signal and background.

3 Statistical methods

The statistical methods used in this thesis are explained in this section. At first a description of NNs is given, then a special type of NN, an AE, is introduced.

3.1 Neural Networks

The basis for NNs were developed by Warren McCulloch and Walter Pitts in 1943 [13], but only became popular in 2011 due to the availability of large parallel computation power in the form of GPUs. The universal approximation theorem states that NNs with one hidden layer and an finite number of hidden nodes can approximate any continuous function on a compact subset of the R^n [24]. The theorem does not state how the parameters have to be tuned to achieve the desired function.

The basic building block of an NN is a neuron that consists of inputs $\vec{I}$, a weight matrix $\mathbf{W}$, a propagation function f , and an activation function φ . An example is shown in Figure 10. The output O of the neuron is given by $O = \varphi(f(\vec{I}, \mathbf{W}))$.

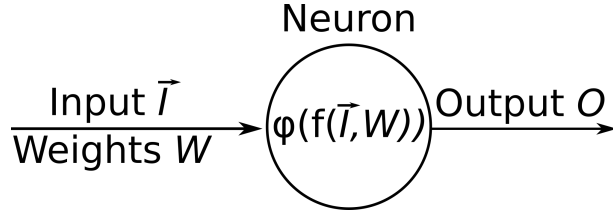


Figure 10: Structure of a neuron where $\vec{I}$ is the input, $\mathbf{W}$ the weight matrix, f the propagation function, φ the activation function, and O the output of the neuron.

The input for the j -th neuron consists of an input vector $\vec{I}$ and a weight vector $\vec{W}_j$, where j is the column index for the weight matrix. The input vector and the weight vector are combined using the propagation function. The most common propagation function is the scalar product of the two vectors

$$f(I, W) = \sum_{i=0}^N I_i \cdot W_i, \quad (8)$$

where I_i is the value of the i -th input value and W_i is i -th weight. In general, a bias unit is included, which is here represented as W_0 where I_0 is set to 1 automatically. A large range of activation functions φ exists. The two most common are the sigmoid function σ and relu (rectifier linear unit), which are defined as

$$\sigma(t) = \frac{1}{1 + e^{-t}} \quad (9)$$

and

$$\text{relu}(x) = \begin{cases} x & x > 0 \\ 0 & \text{else} \end{cases}. \quad (10)$$

An NN consists of layers of neurons, which are divided into three types: the input layer, the hidden layer, and the output layer. While each NN always has one input and one output layer, the number of hidden layers can vary depending on the purpose of the network. The structure of a NN is exemplified in Figure 11.

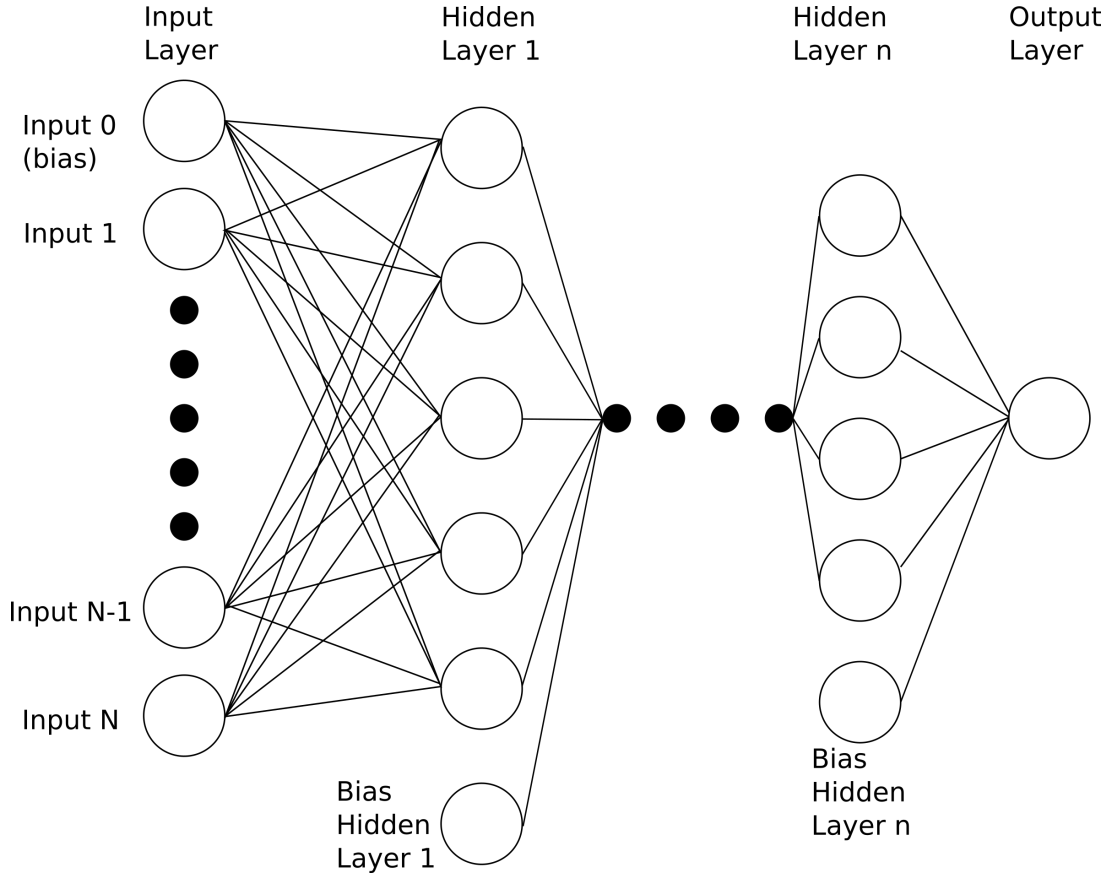


Figure 11: Schematic representation of a NN with n hidden layers. The hidden and output layers take the output of the preceding layer multiplied with the weight matrix as input. The input and each hidden layer contain a bias node.

The training of a NN is done by giving it an input and comparing the output of the NN with the expected output, by using a loss function and then updating the weights accordingly. A common loss for regression problems is the mean square error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (11)$$

where y_i is the expected output for the i -th input vector, $\hat{y}_i$ is the predicted output for the i -th input vector and n is the number of inputs.

The update of the weights is calculated by

$$\Delta w_{ij} = -\eta \frac{\partial MSE}{\partial w_{ij}}, \quad (12)$$

where Δw_{ij} is the update for the j -th weight in the i -th layer, w_{ij} is the j -th weight in the i -th layer, and η is the learning rate.

For deeper hidden layers the weight update is calculated by using the chain rule. When the updates for the weights have been calculated, the weights are updated according to

$$w_{ij}^{new} = w_{ij}^{old} + \Delta w_{ij}, \quad (13)$$

where w_{ij}^{new} is the j -th weight in the i -th layer after the update, w_{ij}^{old} is the j -th weight in the i -th layer before the update. The process of updating the layers from the last to the first on is called backpropagation.

The weights are updated many times in the training process until a satisfactory result is achieved.¹

3.2 Neural Networks for classification

One of the most common use cases for NNs is classification, where a label for each sample exists. This type of method is called supervised learning, because class information for all samples is present. A model for binary classification consists of an input layer, one or multiple hidden layers and an output layer that has a single node. The activation function for the last node is a sigmoid function and the loss function is the binary cross entropy.

¹Other aspects of NN architecture and network training like convolution, dropout, regularization, etc. are not mentioned here to keep this section as concise as possible. If they occur later in the text additional literature will be referenced.

3.3 Autoencoder

An AE is a special type of NN that is mainly defined by its architectural structure. An AE tries to copy the input $\mathbf{x}$ while learning a lower dimensional representation of the inputs. For this task the AE learns two functions, the encoder function $h(\mathbf{x})$ that generates the lower dimensional representation of the inputs and a decoder, $g(h(\mathbf{x}))$ which maps the lower dimensional representation back to the inputs. A general architecture of an AE can be seen in Figure 13. The idea behind using AE for anomaly detection is that the AE is trained only on normal data and learns the latent dimension of the data. If the AE successfully learned an “optimal” representation of the data in the latent dimension and the AE encodes and decodes anomalous samples the reconstruction of the output will have a greater distance to the input than the normal samples. Because the AE is only trained on known data, it has no bias toward a specific anomaly model. This type of model is called semi-supervised because it is only trained on data of one class. The condition, that anomalous inputs have a worse reconstruction than normal inputs, can be expressed as

$$\|g(h(x_{normal})) - x_{normal}\| < \|g(h(x_{anomal})) - x_{anomal}\|, \quad (14)$$

where g is the learned decoder function, h is the learned encoder function, and $x_{normal/anomal}$ is the input.

$\|g(h(x_{normal})) - x_{normal}\|$ is called the reconstruction error Δ_{reco} . When an AE is used for anomaly detection, a threshold for Δ_{reco} is set and all samples, that have a reconstruction error above the threshold, are classified as anomalies. In this work the absolute difference between input and output was chosen. The reconstruction error is given by

$$\Delta_{reco} = |g(h(x)) - x| \quad (15)$$

Figure 12 shows the reconstruction error for the case of data that contains only normal inputs and the case where a small subset consists of anomalous events. The right plot shows a small bump at around 5 which originates from the subset of anomalous inputs.

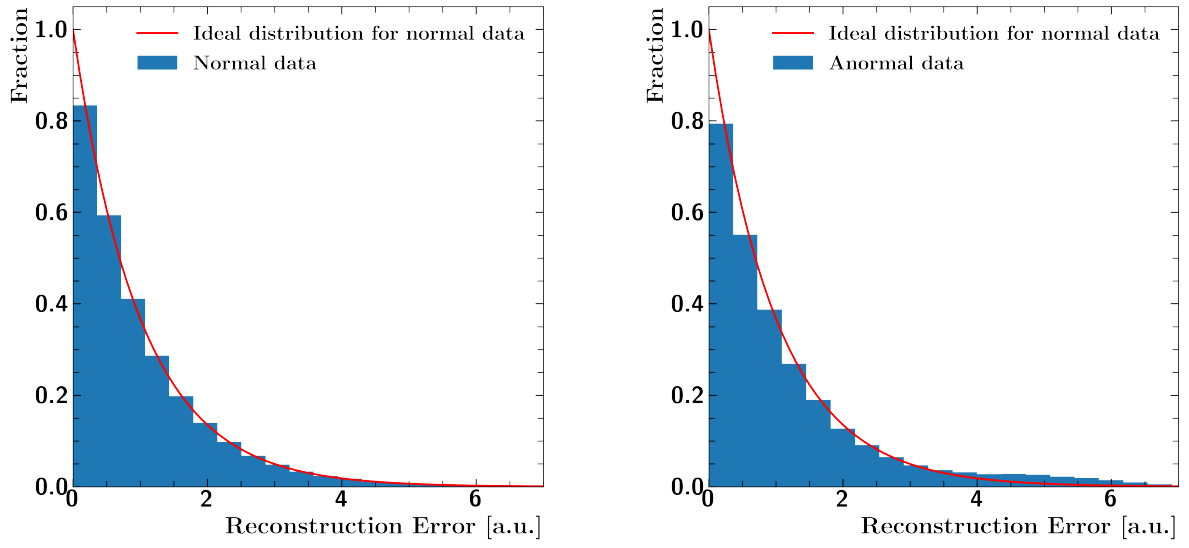


Figure 12: Reconstruction error for a sample of inputs only containing normal data (left) and a sample containing a small subset of anomalous events (right).

Inputs

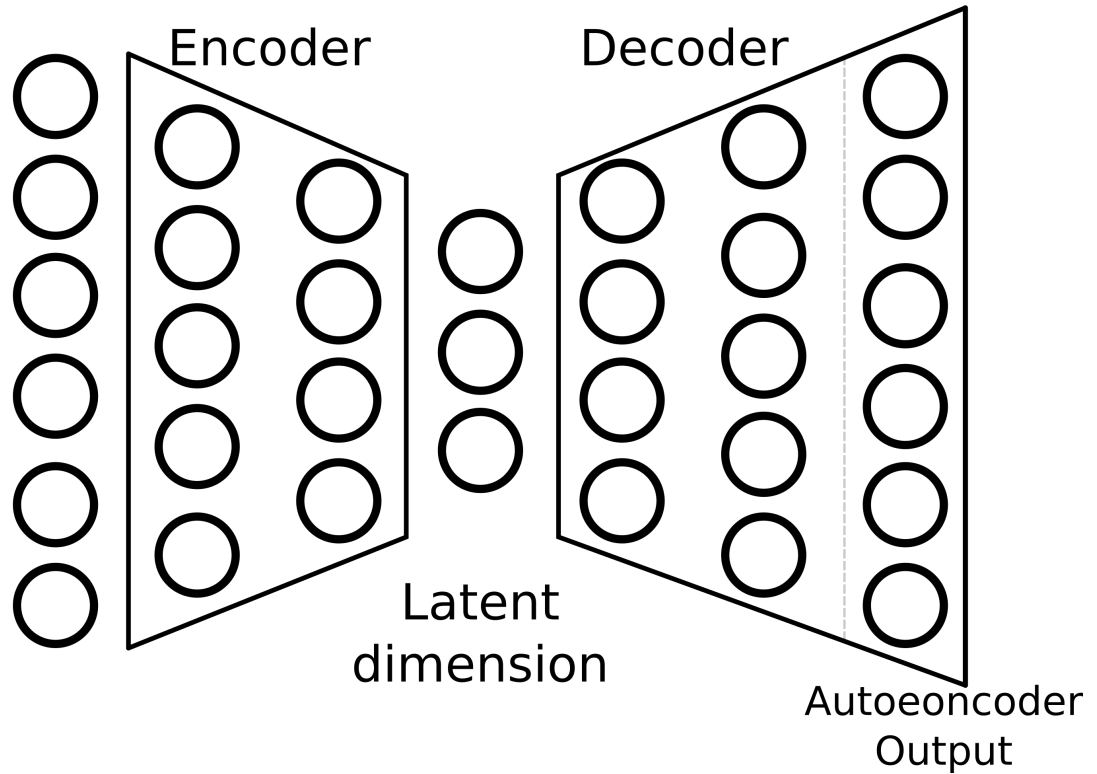


Figure 13: Schematic structure of an AE, where the dimension of the inputs is reduced to the latent dimension with the encoder. The decoder decodes the representation in the latent dimension to be the same as the input.

4 Searches beyond the standard model

Before a detailed analysis based on AEs is performed, a brief overview of different ways to search for new physics is given. Any search looks for an excess above the SM prediction. Searches can be splitted into two groups: model-dependent and model-independent searches. Searches use the same metric for declaring a discovery, the number of events, that exceed the predicted number of events has to exceed the expected yield of events by more than 5σ . The threshold of 5σ is based on the Gaussian probability distribution and is equivalent to a probability of $2.9 \cdot 10^{-7}$ of measuring such an excess by chance if only background events are present.

4.1 Search for new physics with a prior model

When a model for a new particle exists, the decay products and their branching ratios are known to some extent. With that knowledge a region with final states that result from the decay of the new particle can be derived. The selection of the region to search for the particle should have a as high as possible statistic of events, while keeping a high signal-to-background ratio. To optimize the signal-to-background ratio kinematic variables are used and new variables are constructed for better event selection. Cuts are applied on the variables so that a good signal-to-background ratio is achieved, while retaining as many events as possible to have a good statistic. After all cuts are applied the invariant mass of the new particle is reconstructed from the decay products. Figure 14 shows the reconstructed invariant mass of a photon pair originating from a Higgs boson decay.

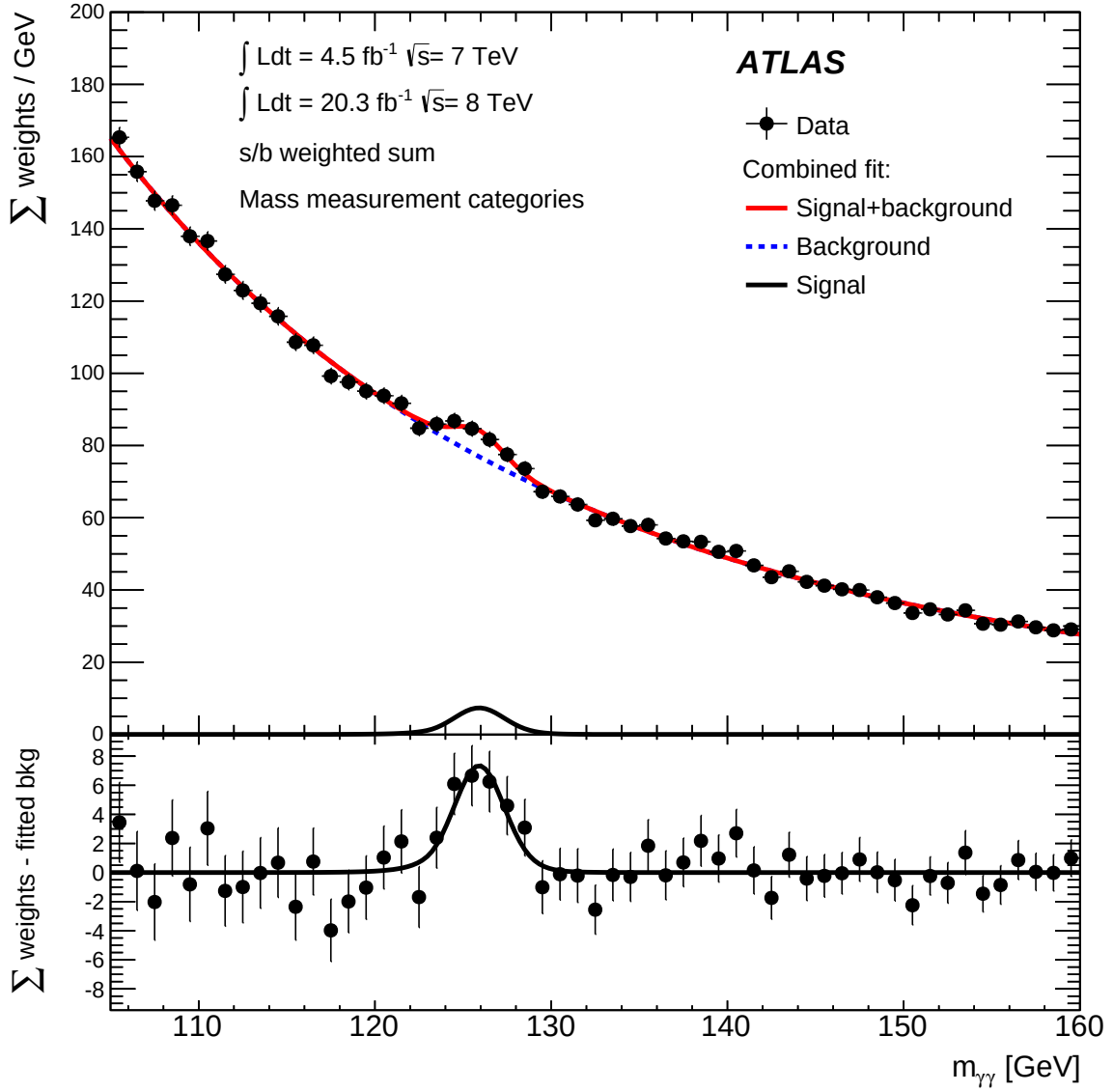


Figure 14: The invariant mass of $\gamma\gamma$ in the upper part. The bump at ≈ 125 GeV comes from a Higgs resonance. In the lower part of the image the fitted background is subtracted from the data [25].

The bump is modeled by the Breit-Wigner-distribution:

$$f(\omega; \omega_0, \gamma) = \frac{1}{(\omega^2 - \omega_0^2)^2 + \gamma^2 \omega_0^2}, \quad (16)$$

where ω_0 is the location of the bump and γ is the width parameter of the distribution.

Another type of searches is looking for new heavy particles like vector-like top quarks. A search of the ATLAS experiment is looking for these particles in the high E_T^{miss} region

due to the fact, that the vector-like top quark can decay into Zt [26]. The Z decays in 20% to a pair of neutrinos [5]. Figure 15 shows a possible production and decay modes for a vector-like top quark. The E_T^{miss} distribution in the signal region of the search is shown in Figure 16. A disadvantage of such searches is that a hypothesis is need such as $T \rightarrow Zt$ and that the analysis and the results are irrelevant when the underlying process is a different one.

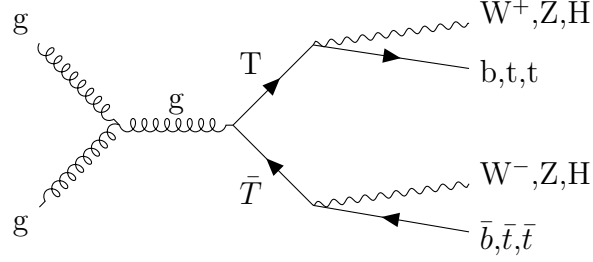


Figure 15: Feynman diagram of the vector-like top quark pair production and possible decay modes.

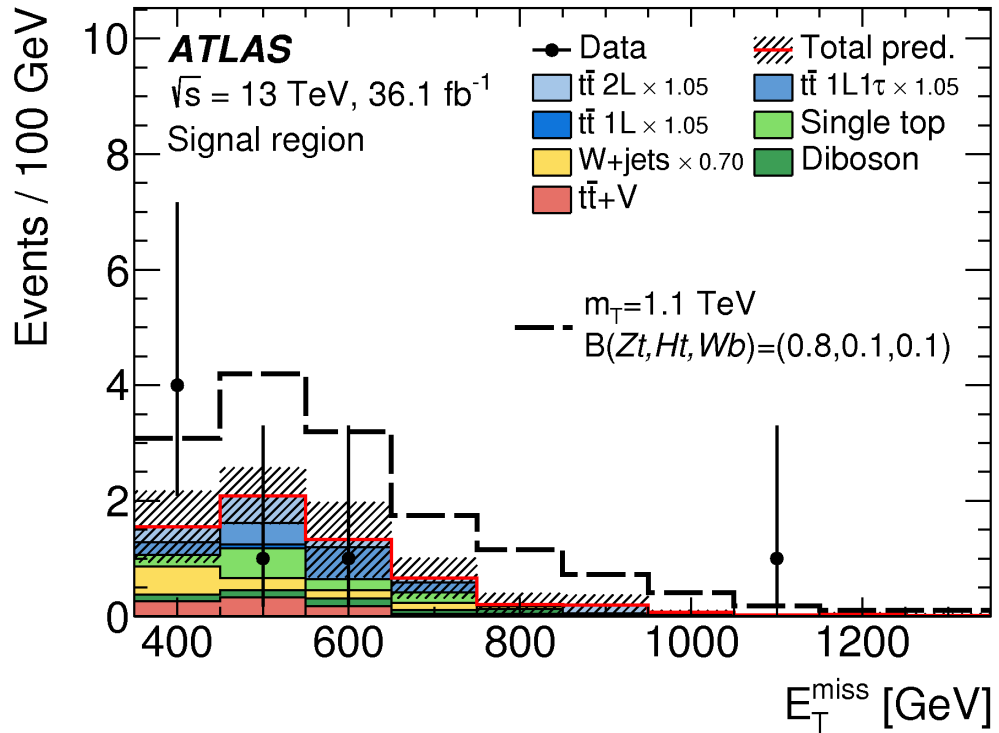


Figure 16: Plot of the E_T^{miss} from a search for vector-like top quarks [13].

In the case of a search for a process, for example flavor-changing neutral-currents (FCNC) [27] the goal is to show that the process generates additional events, that can be

measured in some region. Multivariate methods are used to achieve a better separation between signal and background.

4.2 Search for new physics without a prior model

Searches for hints of new physics without assumptions of the signal have been done before by the DØ Collaboration and CDF Collaboration, both at the Tevatron, by the H1 Collaboration at HERA and ATLAS Collaboration at LHC.

The underlying idea of the search presented in Reference [28] is that the events are split into different categories based on different variables (e.g. type and number of certain objects) and trigger information. For each category a trigger is checked. If the trigger condition is met the event is accepted, else it will either be placed in a different category and a new trigger condition is checked or the event is rejected. An example of the categorization procedure is illustrated in Figure 17. After all events have been categorized, for any category with an excess, additional tests are done. Any region that has an excess at the of p-value 5% or lower level is further investigated by a more thorough analysis. The paper reports no findings.

4.3 Using Autoencoders for new physics searches

AEs are suited for searches with no prior model of the signal. They only need simulated MC events of SM processes for training. After the training the reconstruction error is calculated for the samples the AE was trained on and measured data. The distribution of the reconstruction error is compared between the measured data and the SM background. A significant deviation in this distribution could be interpreted as evidence for physics beyond the SM. If a new signal is detected by the AE at some threshold, the ratio between the reconstruction error of the MC simulation and the measured data increases. If no new signal can be detected, the ratio between the MC simulation and data should be uniform or mimic the structure of the ratio present in the kinematic variables. In the general search described above AEs would be used for each category to automatically look for a new signal in that region. A further advantage of an AE is that it can be used to get a better understanding where to look for a new signal or how the new signal differs from the background. The cost of using an AE is very low due to the only requirement being MC samples of the relevant SM processes. All other steps of the tuning and variable selection for the AE can be done automatically.

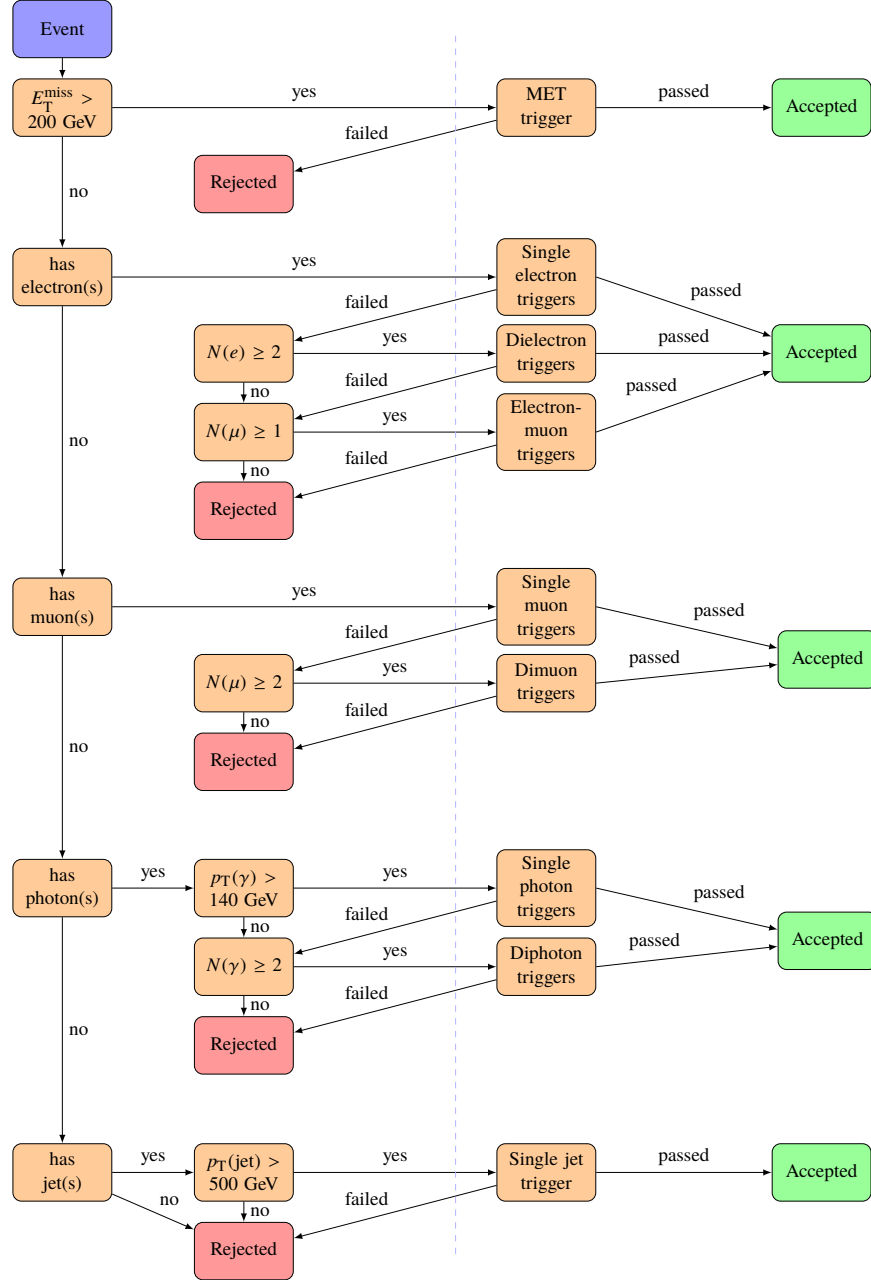


Figure 17: The figure shows an example of the selection process detailed above [28].

5 Analysis of Autoencoders under controlled conditions

In this section a test of AEs under controlled conditions is described. The goal is to train AEs on $t\bar{t}$ events and find the tW process present in the data. The performance of the AE is compared against a physically motivated variable named mini-max $m(b\bar{l})$ [29] and a supervised NN trained for classification.

5.1 Motivation

The pair production of top quarks is one of the most studied processes at the LHC and also the major background to most BSM searches. One process that interferes with the pair production of $t\bar{t}$, is the production of a single top quark in association with a W boson and a b quark. This work uses a selection of $2l2b$ ($1l$ has to be an electron and the other a muon) in the final state, which yields a sample that mainly contains $t\bar{t}$ and single top events with minor contributions from fake-lepton background and $t\bar{t}V$ [29]. Fakes are leptons that do not originate directly from a W or Z decay but from some other source like photon conversion or a misclassification of a hadron as a lepton. The $t\bar{t}V$ process describes $t\bar{t}$ production with the radiation of a Z or W boson.

5.2 Data and simulated samples

This section is based on the information given in Ref. [30].

In this work the full Run 2 dataset collected by ATLAS at LHC during 2015-2018 at a $\sqrt{s}=13$ TeV with a total integrated luminosity of 139 fb^{-1} is used. The detector and trigger simulations were done with GEANT4 [31, 32]. For the simulation the full ATLAS detector simulation was used. Table 1 gives an overview of all simulated samples at the end of the section.

$t\bar{t}$

This section describes the production of the $t\bar{t}$ MC sample. For the production of $t\bar{t}$ events POWHEGBOX v2 [33, 34, 35] was used, it provides MEs at next-to-leading order (NLO) in the strong coupling constant α_s with NNPDF3.0nlo for the PDFs [36] and the h_{damp} parameter was set to $1.5 m_{\text{top}}$ [37]. For renormalization and factorization, the scale of the functional form was set to the default scale. The decay of top quarks was simulated at LO with MADSPIN [38, 39]. For parton shower and hadronization PYTHIA8 (v8.230) [40], with the A14 set of tuned parameters [41] and NNPDF3.3lo. For the decay of bottom

and charm hadrons `EVTGEN v1.6.0` [42] was used.

tW

The events were generated with `POWHEGBOX v2`, it provides MEs at NLO in the five flavor scheme with `NNPDF3.0n1o` PDF set. The functional form of the factorization and renormalization scale was set to the default value of m_{top} . To handle the interference with $t\bar{t}$ production the diagram removal scheme [43] was used. The decay of top quarks was simulated at LO with `MADSPIN`. For interfacing the events `PYTHIA8 v8.230` with, the `A14` tune and the `NNPDF2.31o` PDF set was used. For the decay of bottom and charm hadrons `EVTGEN v1.6.0` [42] was used.

Fakes

Fakes are events, where at least one of the leptons does not originate from a Z or W boson decay, but from the decay of a bottom or charm hadron, a photon conversion to an electron, the decay of a pion or kaon to a muon, or a misidentified hadron. Fakes are estimated with a partially data-driven technique. It used the $e\mu$ same sign (SS) control region, because most of the events in that region are fakes.

$t\bar{t}V$

The production of $t\bar{t}V$ is modeled with the `MADGRAPH5_aMC@NLO v2.3.3` [44] generator at NLO using the `NNPDF3.0n1o` [36] PDF. The events are interfaced to `PYTHIA8 (v8.128)` [40] using the `A14` tune [41] and the `NNPDF2.31o` [36] PDF set. Decays of bottom and charm hadrons are simulated with `EVTGEN v1.2.0` [42].

Table 1: Overview of the simulated samples.

Process	ME Generator	ME PDF	Parton shower/ hadronization
$t\bar{t}$	<code>POWHEGBOX</code>	<code>NNPDF3.0n1o</code>	<code>PYTHIA8</code>
tW	<code>POWHEGBOX</code>	<code>NNPDF3.0n1o</code>	<code>PYTHIA8</code>
$t\bar{t}V$	<code>MADGRAPH5_aMC@NLO</code>	<code>NNPDF3.0n1o</code>	<code>PYTHIA8</code>

5.3 Object definition and event selection

This section is based on the information given in Ref. [30]. In this section the definitions of the objects used in this analysis are given.

5.3.1 Object definition

Electrons

Electrons are identified from the tracking information of the ID and energy deposition in the electromagnetic calorimeter. Electron candidates have to fulfill the TightLH quality requirements, which combine a multivariate likelihood fit with a series of tracking and calorimetric variables. Furthermore, the electron candidates are required to fulfill the Gradient isolation requirements. The final selection is based on p_T , each candidate needs to have a $p_T > 25$ GeV and the calorimeter cluster position $|\eta_{cl}|$ has to be below 2.47 with an exclusion of the transition region between 1.37 and 1.52.

Muons

Muons are identified by tracking information from the muon detectors and the ID. Candidates have to pass the medium quality requirement. The muons also have to fulfill the Gradient isolation requirement. The transverse momentum has to be above 25 GeV and $|\eta|$ below 2.5.

Jets

Jet candidates are reconstructed from topological clusters [45], with the anti- k_t algorithm [46] and a radius requirement of $R=0.4$. Reconstructed jets have to fulfill $p_T > 25$ GeV and $|\eta| < 2.5$. Furthermore, to suppress jets from pileup, the jet-vertex-tagger (JVT) [47] is used. The jet candidates are required to have a JVT value of 0.59 or more if $p_T < 120$ GeV and $|\eta| < 2.4$.

b-jets

b-jets are identified by algorithms [48], that exploit the long lifetime, high mass, high decay multiplicity of b-hadrons, and properties of the b-quark fragmentation. These algorithms use a two-stage approach, where in the first step low level algorithm reconstruct characteristic features of the b-jet and in second step high level multivariate classification algorithms are used. This analysis uses the MV2 algorithm [48], that is based on a Boosted Decision Tree (BDT). The performance of MV2 is evaluated using a b-jet tagging single cut operation point (OP). The OP is based on a selection on the MV2 discriminate distribution and ensure a specific b-jet efficiency ϵ_b . The b-jet efficiency was set to 70%.

Overlap removal

To avoid double-counting reconstructed objects, the following overlap removal steps were done.

Electrons: Most of the time electrons are also reconstructed as jets. The jet candidate is thrown away when it matches an electron within $\Delta R < 0.2$. If any remaining electron candidate is still matched with a jet candidate within $\Delta R < 0.4$ the electron candidate is thrown away.

Muons: Muons inside jets originate in many cases from semi-leptonic heavy-flavor decays. Thus, muon candidates are removed if these overlap with a jet of $p_T < 25$ GeV within $\Delta R < 0.4$ and which passed the JVT cut.

Electrons and muons: If a reconstructed electron and muon share a track the electron candidate is thrown away.

5.3.2 Event Selection

Recorded data events are only selected if all sub-detectors were working properly. The information for this is stored in a Good Run List (GRL). Furthermore, events are removed if a noise burst occurred in the calorimeter. All other requirements are set on the reconstructed objects. All events are required to have at least one primary vertex, that is reconstructed from at least 5 tracks. The events are rejected if they contain a jet that does not pass the loose jet cleaning criteria, or a badly measured muon. Furthermore, all events have to pass at least one single lepton trigger. Events are required to have two or three jets with $p_T > 25$ GeV. Of the jets exactly two have to be b-tagged jets with a working point of 70%. Furthermore, two leptons with one being an electron and the other an opposite charge muon are required. The electron needs to have $p_T > 25$ GeV and the muon $p_T > 25$ GeV, additionally for events from 2016-2018 one of the leptons are required to have a $p_T > 27$ GeV. The invariant mass of the two leptons $m(e, \mu)$ has to be above 15 GeV. No requirement for E_T^{miss} was made.

The number of expected events from simulation and observed events after the selection is shown in Table 2. When only accounting for the $t\bar{t}$ process, the fraction of data that is not described by the $t\bar{t}$ events is given by $1 - \frac{Events_{t\bar{t}}}{Events_{data}} = 5.4\%$.

5.4 Kinematic properties

In this section the distribution of kinematic variables are presented, to get a better understanding of them and to better understand how the algorithms select events. Plots for all 26 variables that were used as input for the AE were generated. To keep this section

Table 2: Observed and expected events after the selection.

Process	Events
$t\bar{t}$	165995
tW	5538
fakes	1340
$t\bar{t}V$	202
sum	173075
data	175460

concise only a selection of 8 plots is given. All other plots can be found in Appendix A.1.

Figure 18 shows p_T and $|\eta|$ for the electron and primary b-jet (b-jet with the highest p_T). The distribution of $p_T(e)$ has its peak around 25 GeV, the peak for $p_T(b1)$ is around 80 GeV with a steep slope to the left and a bit flatter slope to the right. For the electron the slope of $|\eta|$ is steeper than for the primary b-jet. The gap in the plot for the electron originates from the exclusion of that region.

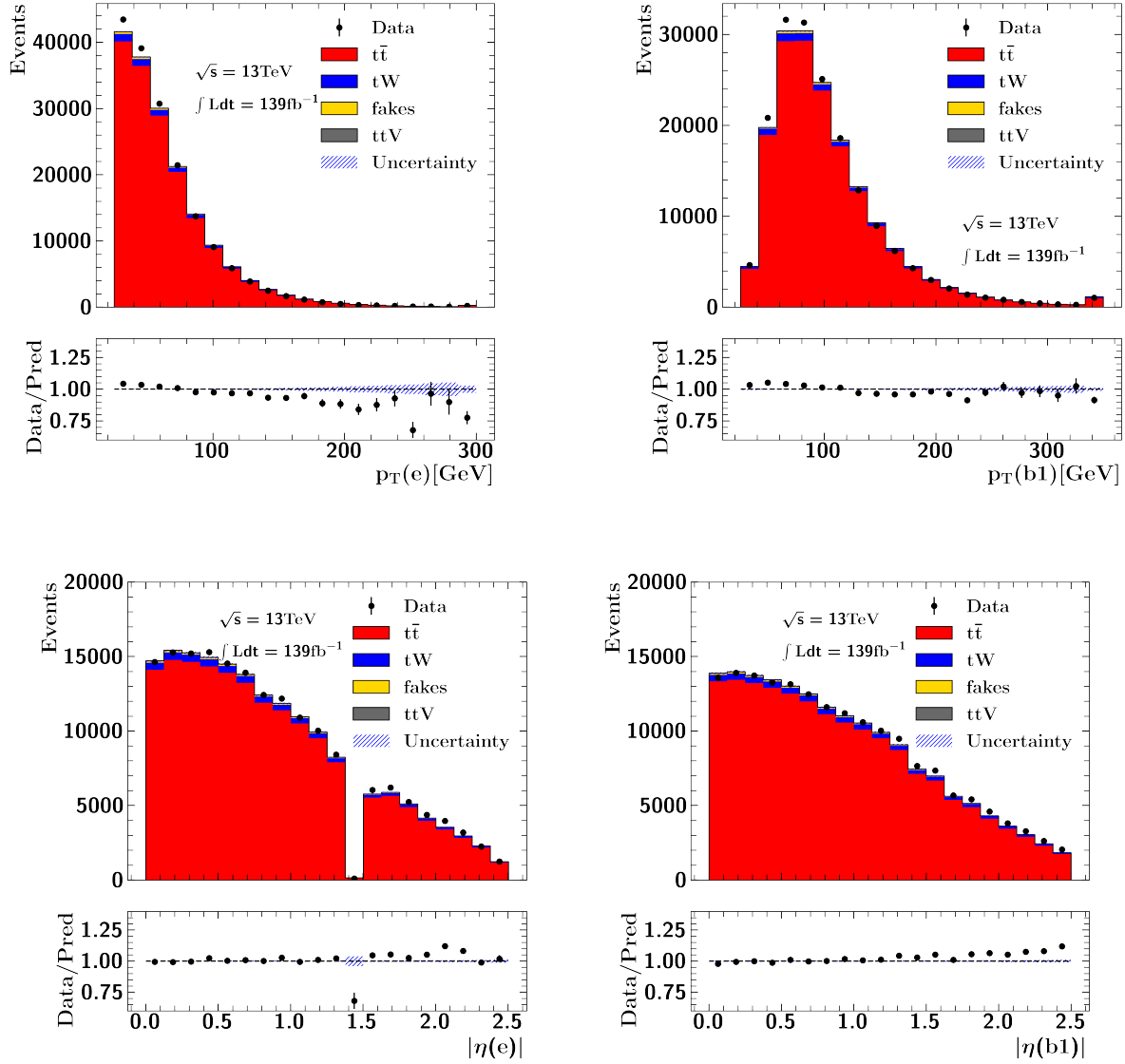
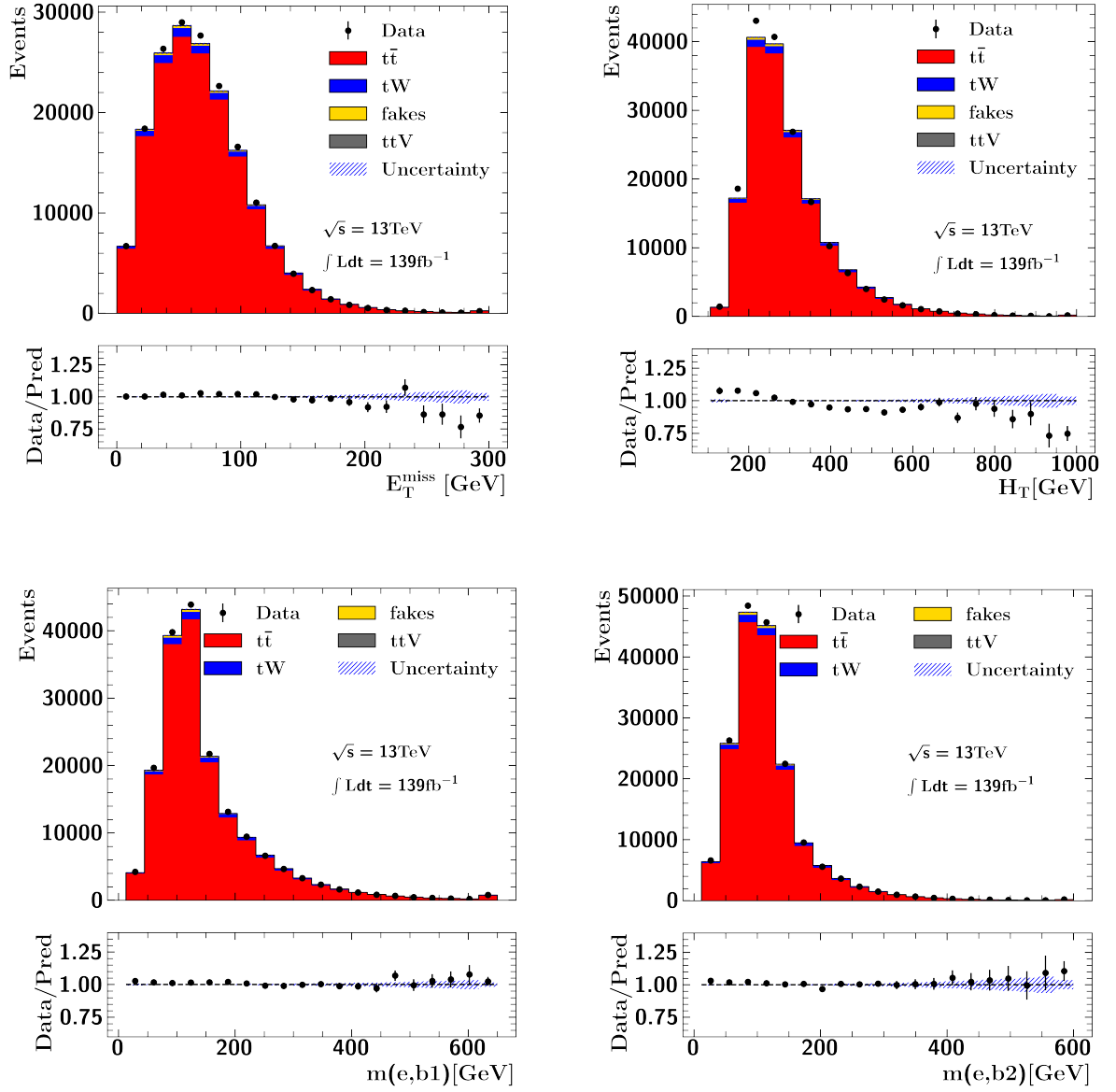


Figure 18: The plot shows p_T and $|\eta|$ for the electron and the primary b-jet.

Figure 19 shows the distribution of E_T^{miss} , H_T , $m(e, b1)$, and $m(e, b2)$. H_T is the sum over the transverse momenta of all reconstructed objects and E_T^{miss} . $m(e, b1)$ and $m(e, b2)$ are the invariant mass of the reconstructed electron and the leading/sub-leading jet respectively. The peak of the E_T^{miss} distribution is at around 50 GeV. H_T is centered around 250 GeV and has a long tail towards higher values. The distributions of $m(e, b1)$ and $m(e, b2)$ peak around 150 GeV.

Figure 19: The plot shows E_T^{miss} , H_T , $m(e, b1)$ and $m(e, b2)$.

5.5 Analysis strategy

To compare the three different approaches of mini-max m(bl), the AE, and the NN classifier a metric is needed, that is not favorable to any of the algorithms. As performance metric the discrepancy between data and the expected $t\bar{t}$ events is used. The discrepancy D is defined as

$$D = 1 - \frac{Events_{t\bar{t}}}{Events_{data}}. \quad (17)$$

This gives the relative number of events that are not described by the $t\bar{t}$ process. When looking at the data without any additional selections or cut beyond the event selection described in Section 5.3. The discrepancy is 5.4%, denoted in the following as the baseline discrepancy.

The order of presenting the algorithms (mini-max m(bl), AE, NN classifier) was chosen this way to give the reader a better physical understanding of the processes involved. The actual order of the algorithms when conducting the experiment was AE, mini-max m(bl), NN classifier. For the training of the AE only the simulated $t\bar{t}$ events were used. Only after the final model of the AE was selected and a cut on the reconstruction error was chosen, the data and the simulations of the other processes were used. This was done to ensure no biases during the development of the AE model. For the final AE a cut on the reconstruction error was made. After the cut 5386 event of data were left. To make the results comparable at the same signal yield the cut on the discriminates were set to yield close to 5300 events for the measurement data.

5.6 mini-max $m(\text{bl})$

Requiring that the final state has 2 b-jets and 2 leptons ensures a rather pure selection of $t\bar{t}$ and tW events. The difference between the $t\bar{t}$ process and the single top process lies in the resonant structure of the Wb pairs. In the $t\bar{t}$ case the invariant mass of the W and the corresponding b is $m^2(W, b) \sim m_{top}^2$. In the single top process one of the Wb pairs is off the top mass shell. To exploit this fact, a variable called mini-max $m(\text{bl})$ is constructed to discriminate between $t\bar{t}$ and the tW events.

For the b and l that originate from on-shell produced top quark $m(b, l) \leq \sqrt{m_{top}^2 - m_W^2}$ holds, but not for the off-shell b, l -pair. Having two b-jets and two leptons, labeled as b_1, b_2, l_1 and l_2 , four invariant masses $m_{b_1, l_1}, m_{b_1, l_2}, m_{b_2, l_1}, m_{b_2, l_2}$ can be constructed. For the $t\bar{t}$ process two of the combinations have to be below the top mass bound. For the single top process only one of the b, l -pairs has to be below the top mass bound. The mini-max $m(\text{bl})$ variable takes advantage of this difference and is defined as

$$\mathbf{mini-max} \, m(b_{1,2}, l_{1,2}) = \min(\max(m_{b_1, l_1}, m_{b_2, l_2}), \max(m_{b_1, l_2}, m_{b_2, l_1})). \quad (18)$$

$\min$ selects the smallest of the arguments and $\max$ selects the biggest. First the maximum of the b, l -pairs is taken to take advantage of the fact, that for the $t\bar{t}$ process $m(b, l) \leq \sqrt{m_{top}^2 - m_W^2}$ applies. Then to ensure the b, l -pairs were chosen correctly the minimum of the two resulting pairs is taken. With this construction events that result from the $t\bar{t}$ process will have mini-max $m(\text{bl})$ values below the top mass bound and events that result from the single top process will have values over the whole range, especially above the top mass bound [29].

Figure 20 shows the distribution of mini-max $m(\text{bl})$. On the left, only data and the $t\bar{t}$ process is shown, whereas on the right the other MC samples are included. The ratio between data and predicted $t\bar{t}$ events (lower part of the plots) show an increase on the left around the top quark mass. The increase mainly comes from the off-shell b, l -pairs and is described by the other MC events.

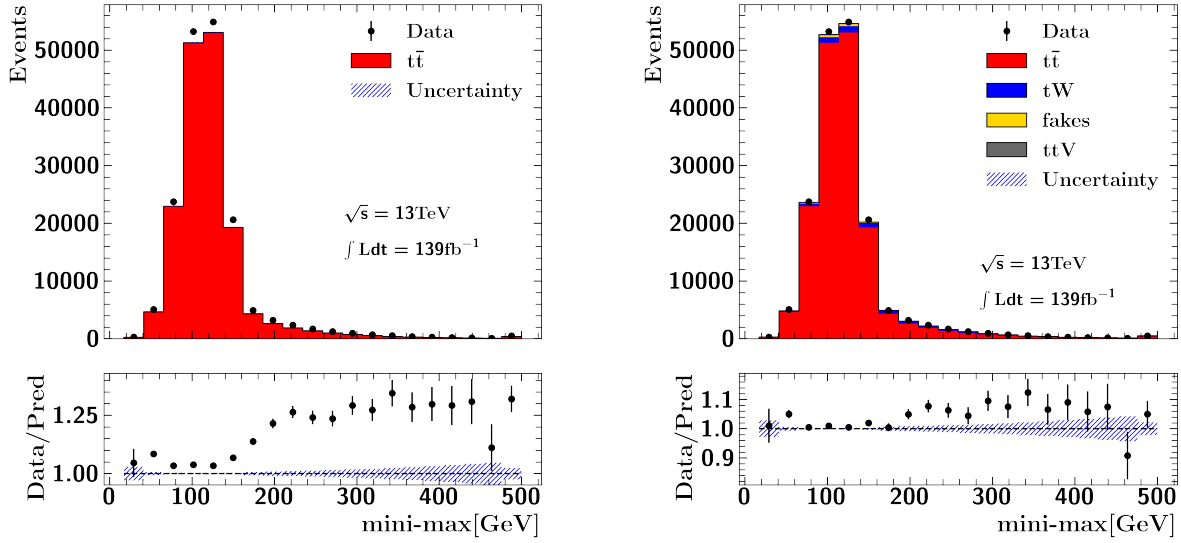


Figure 20: mini-max $m(\text{bl})$ variable with data and only simulated $t\bar{t}$ events (left) and data and all simulated samples (right). In the plot on the left an increase in the ratio can be seen around the top mass. The discrepancy is not appearing as a step because of measurement uncertainties. The residual discrepancy that can be seen on the right is expected, because the MC simulations for $t\bar{t}$ and tW do not include interference of higher amplitudes.

The cut on mini-max $m(\text{bl})$ was set to 260 GeV which has the event yield shown in Table 3. The resulting discrepancy is 21.9% which is an increase of 16.5% compared to the 5.4% baseline discrepancy.

Table 3: Event yield after introducing the cut on mini-max $m(\text{bl})$ of 260 GeV.

Process	Events
data	5220
$t\bar{t}$	4072
tW	753
fakes	33
$t\bar{t}V$	19

5.6.1 Variables after mini-max $m(b\ell)$ cut

To get a better understanding of the events, that were removed by the cut on mini-max $m(b\ell)$ this section shows some of the kinematic variables after the cut of mini-max $m(b\ell)$.

Figure 21 shows the distributions of p_T and η for the electron and the primary b-jet after the cut. Compared to the plots in Figure 18 mainly low p_T events were removed by the cut. For $|\eta|$ most or the removed events have a low $|\eta|$. This is specially apparent in the $|\eta(b1)|$ plot (lower right).

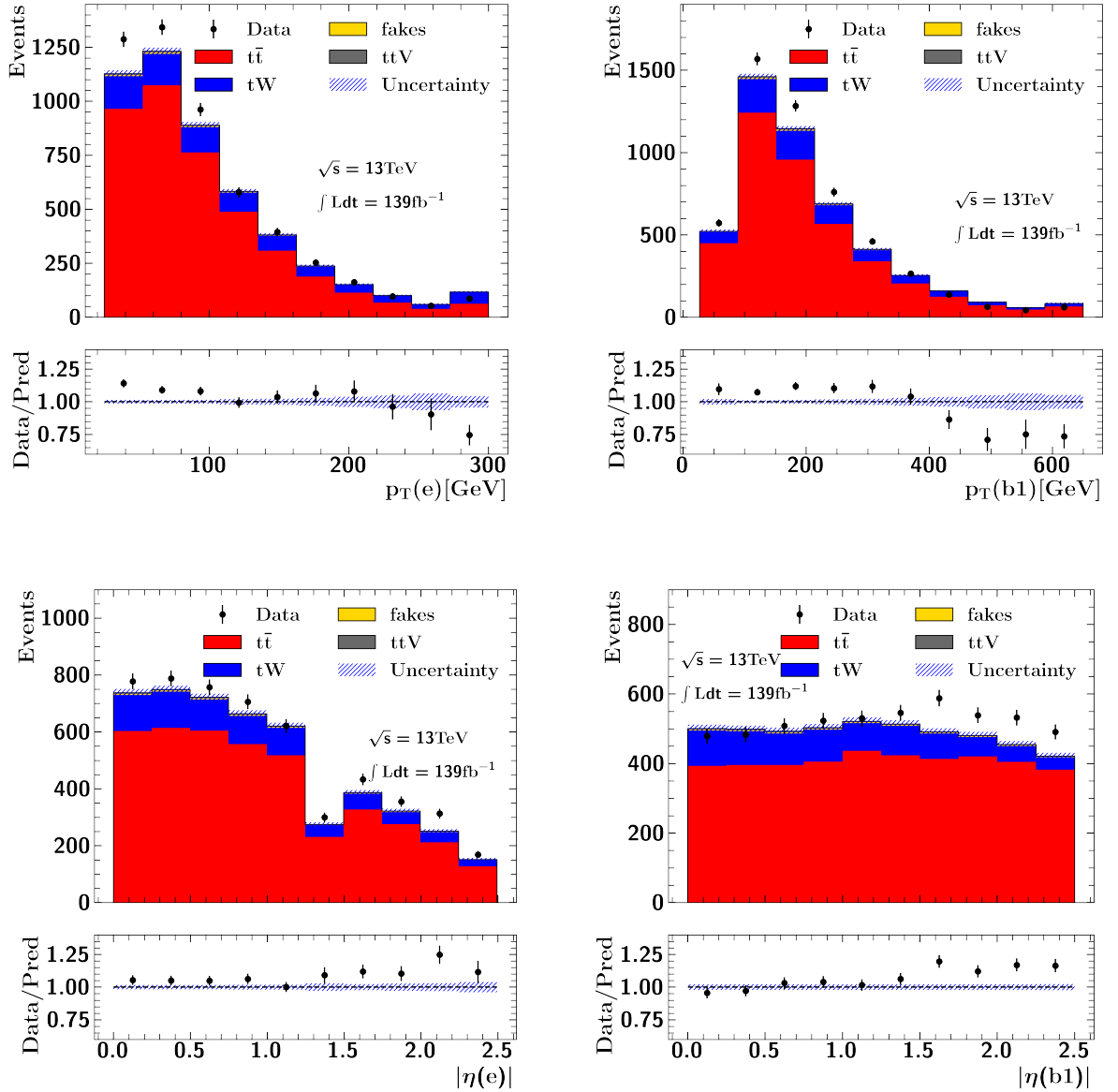


Figure 21: Plots of $p_T(e, b1)$ and $|\eta(e, b1)|$ after the cut on mini-max $m(b\ell)$.

Figure 22 shows the distributions of E_T^{miss} , H_T and $m(e, b_{1,2})$. For E_T^{miss} and H_T the main fraction of events are in the lower region of the two variables. The distributions

of $m(e, b_{1,2})$ changed from unimodal distributions to bimodal distributions, this is to be expected from the construction of mini-max $m(\text{bl})$. The first peak comes from the on-shell produced Wb pair and the second peak comes from the off-shell produced pair.

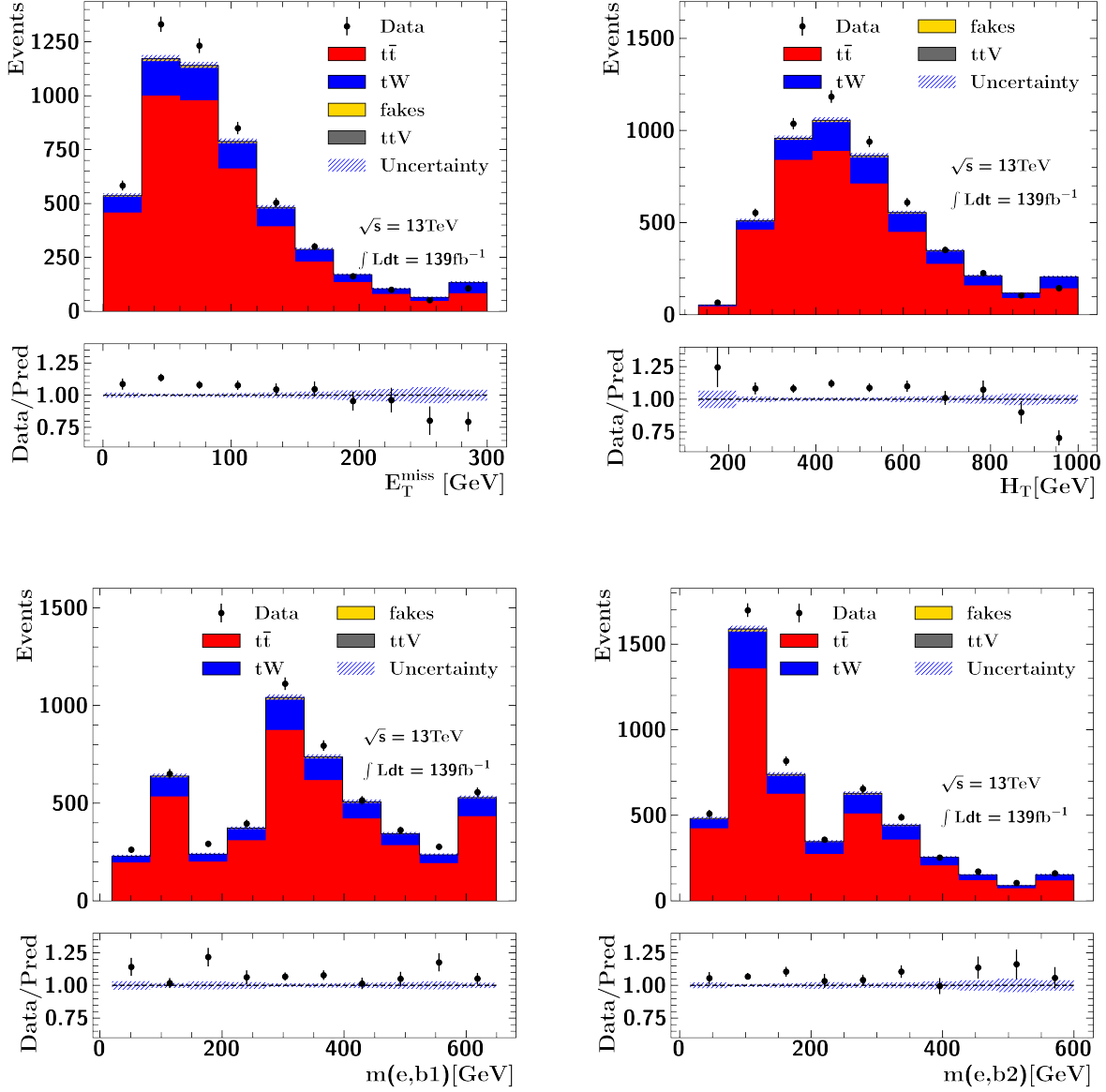


Figure 22: Plots of E_T^{miss} , H_T , $m(e, b1)$ and $m(e, b2)$ after the cut on mini-max $m(\text{bl})$.

5.7 Autoencoder

This section describes the usage of an AE for anomaly detection and is split into three parts. The first part describes the preprocessing of the data, the second part explains the training and architecture of the AE, and the last part presents the results.

5.7.1 Variables and preprocessing

Due to the huge differences in the range of the variables: $|\eta|$ from 0 to 2.5 or H_T from 100 GeV with a tail above 2000 GeV, a preprocessing has to be done to ensure a smooth error surface. For the input different sets of variables were tested the best results were produced with a set of 26 variables. The variables are listed in Table 4.

Table 4: List of variables that were used as input for the AE.

$p_T(e)$	$E(e)$	$ \eta(e) $	$\phi(e)$
$p_T(\mu)$	$E(\mu)$	$ \eta(\mu) $	$\phi(\mu)$
$p_T(b1)$	$E(b1)$	$ \eta(b1) $	$\phi(b1)$
$p_T(b2)$	$E(b2)$	$ \eta(b2) $	$\phi(b2)$
$M(e,b1)$	$M(\mu,b2)$	$M(e,b2)$	$M(\mu,b1)$
$\Delta R(e, \mu)$	$\Delta R(l1, b1)$	$\Delta R(l2, b2)$	
E_T^{miss}	$\phi(E_T^{miss})$	H_T	

As preprocessing standardization, power transformations, and scaling from 0 to 1 were tested. Standardization produced bad results for all log normal distributed variables. Alternatively, power transformations [49] were tested but still yielded bad results. For that reason, only scaling of all variables from 0 to 1 was used as preprocessing.

5.7.2 Architecture

The AE uses a symmetric structure around the latent dimension, with 9 layers in total. The input layer has a size of 26 nodes and the latent dimension has a dimension of 16. The compression in the latent dimension is $16/26 = 61.54\%$. For the first 3 layers the number of nodes is reduced by 3 per layer and the fourth layer reduces the number of nodes by 1, which yields a structure of $26 \rightarrow 23 \rightarrow 20 \rightarrow 17 \rightarrow 16 \rightarrow 17 \rightarrow 20 \rightarrow 23 \rightarrow 26$ nodes per layer. This structure was found to produce the best results after testing different number of layers and number of nodes reduced per layer. For the different hyperparameters of the AE, a hyperparameter study was done. Table 5 shows the hyperparameters and the values that were tested for them. The training took about 50 mins with 3.6 million inputs. To

prevent over fitting early stopping was used for the validation loss with a patience of 15. Early stopping stops the training when a specified value (here the validation loss) does not improve for a defined number of epochs. The patience parameter defines for how many epochs the specified values must not improve, before the training is stopped.

Table 5: Hyperparameters that were searched when optimizing the AE. Final set of parameters is given below in Table 6

Parameter	Values
Activation function	sigmoid, relu, leaky relu, selu [50]
Learning rate	0.001, 0.0001, 0.00001
Optimizer	SGD, Adam [51]
Loss function	MAE, MSE
Architecture	see above

Overall, 60 different hyperparameter combinations have been tested for the final set of layers. Table 6 shows the final set of hyperparameters that produced the best results from the hyperparameter search.

Table 6: Final set of hyperparameters for the AE, that produced the best results.

Parameter	Values
Activation function	selu
Learning rate	0.0001
Optimizer	Adam
Loss function	MSE
Architecture	see above

The AE was built with Tensorflow [52], a machine learning framework, with extended capabilities for deep learning.

5.7.3 Results

In this section the results of the AE are discussed. The AE was trained using only simulated $t\bar{t}$ events as input. Figure 23 shows the reconstruction error for the trained AE, the threshold was set to 0.3. The upper right and lower plot show the reconstruction error with different sets of simulated samples. The plot in the top left shows only the $t\bar{t}$ events and data, at a reconstruction error of 0.3 an increase of the discrepancy between the MC sample and data can be observed. The plot in the top right shows, that the discrepancy mainly comes from the missing tW process. In the lower center a plot including all considered processes shows that the effect of $t\bar{t}V$ and fakes is only minor. After the cut the discrepancy between $t\bar{t}$ events and data is 16.52% which is an increase of 11.14% from the baseline. The number of events after the cut for each process are shown in Table 7.

Table 7: Number of events after the cut on the reconstruction error, for each process.

Process	Events
data	5386
$t\bar{t}$	4496
tW	537
fakes	36
$t\bar{t}V$	18

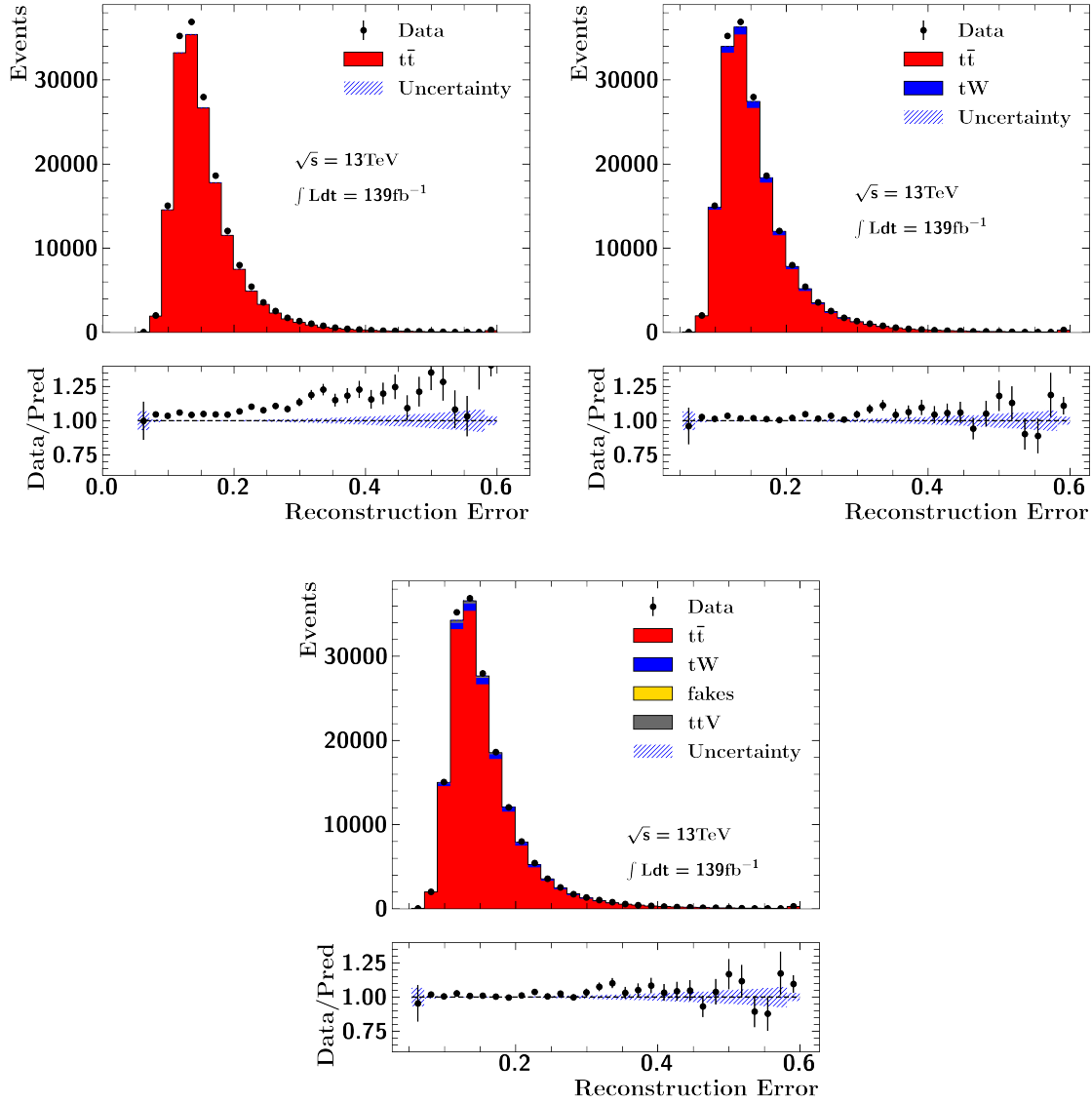


Figure 23: The reconstruction error (Equation 15) with different sets of MC samples and data is shown. The AE was only trained on the $t\bar{t}$ MC sample. The plot on the top left shows the reconstruction error for the $t\bar{t}$ MC sample and data, above the threshold of 0.3 the discrepancy increases. The plot in the top right shows the reconstruction error for the $t\bar{t}$ MC sample, data, and the tW process. The tW MC sample accounts for most of the discrepancy.

Figure 24 shows the distribution of p_T and $|\eta|$ for the electron and the primary jet after the cut. For p_T the biggest fraction of removed events comes from the low p_T region. When looking at $|\eta|$, most of the events removed by the AE have a low $|\eta|$. This removal is stronger as the removal by mini-max m(bl), as can be seen in comparison to Figure 21.

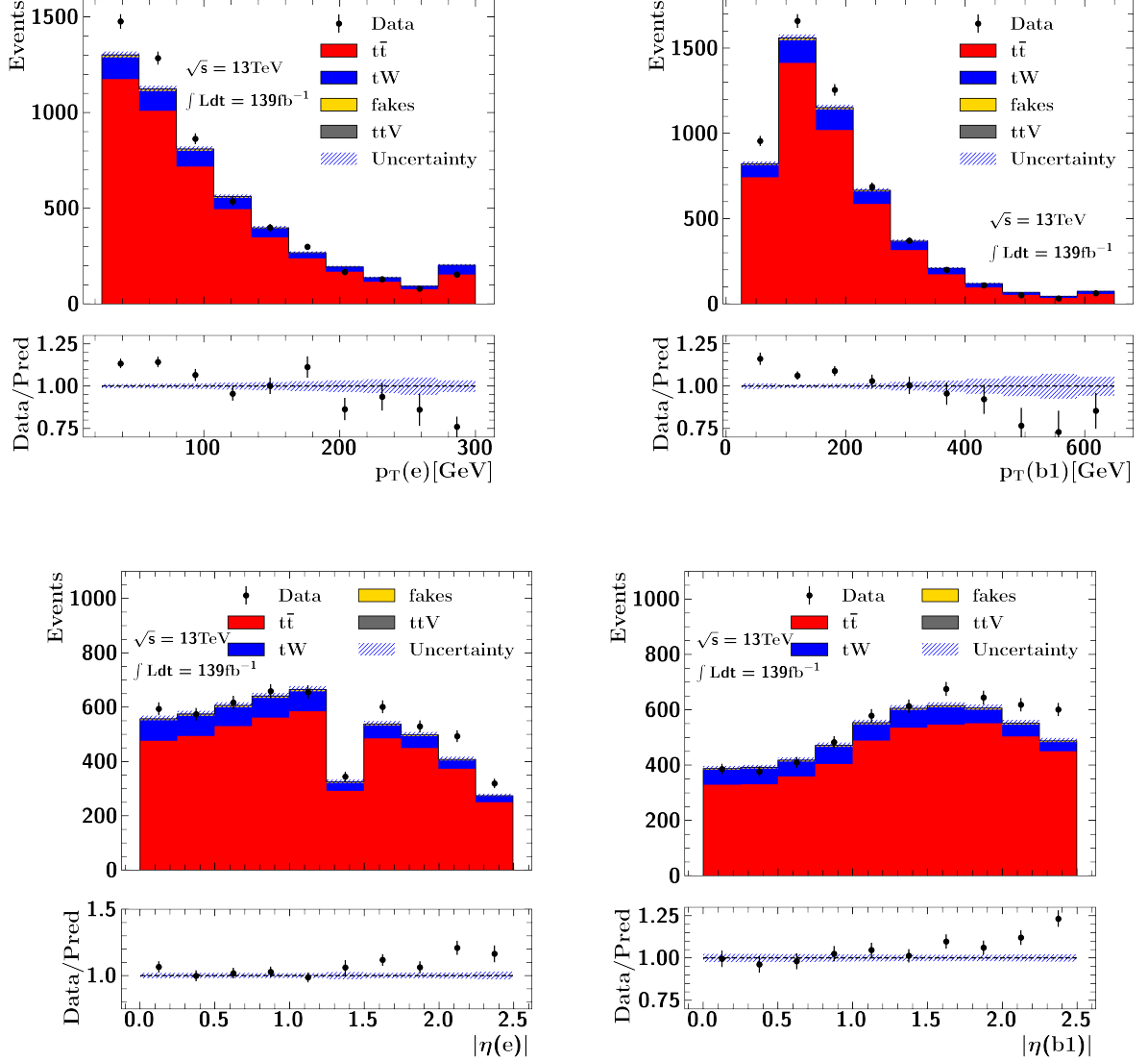


Figure 24: Plots of $p_T(e, b1)$ and $|\eta(e, b1)|$ after the cut of the AE has been applied.

Figure 25 shows the distributions of E_T^{miss} , H_T , $m(e, b1)$, and $m(e, b2)$. Most of the removed events for E_T^{miss} and H_T come from the lower region, the shapes of the distributions is quite similar to the ones for mini-max m(bl) shown in Figure 22. For $m(e, b_{1,2})$ the distribution after the cut of the AE are missing the strong bimodal peaks, that can be seen in Figure 22 (lower two plots). The first peak from the on-shell top quark is present but the second peak from the off-shell b,l-pair is much smaller. This leads to the conclusion that the AE has learned an internal representation of mini-max m(bl) but also uses other

properties to determine if an event is an anomaly or not.

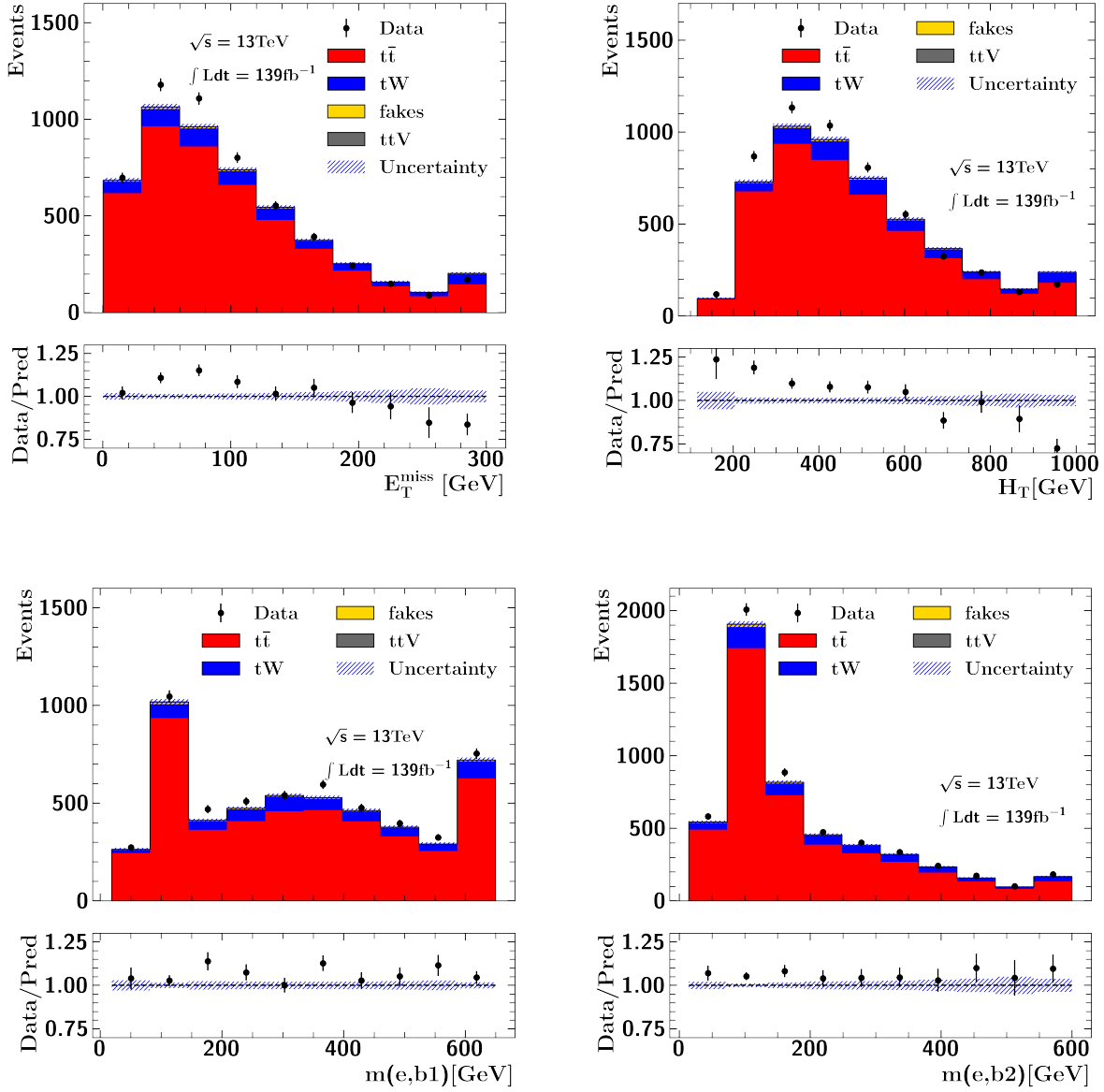


Figure 25: Plots of E_T^{miss} , H_T , $m(e,b1)$, and $m(e,b2)$ after the cut of the AE has been applied.

5.8 Classifier

In this section an analysis with a NN classifier is done to compare the results of an unsupervised algorithm (AE) against a fully supervised algorithm (NN classifier).

5.8.1 Preprocessing

To make the classifier comparable to the AE, the variables and the preprocessing are the same as described above in Section 5.7.1.

5.8.2 Architecture

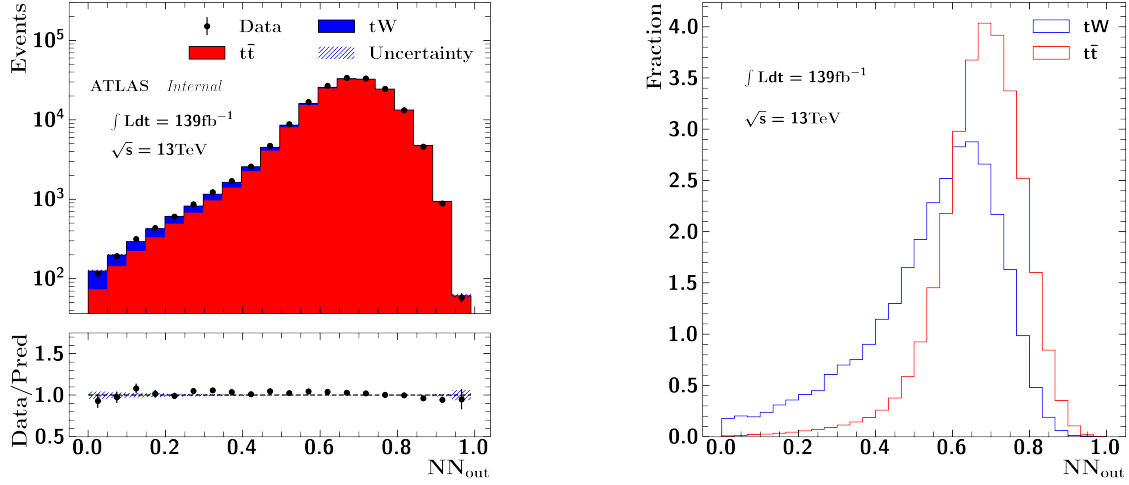
The classifier has two hidden layers with 128 nodes each and a single output node. The activation function for the hidden nodes is selu and for the output node sigmoid. To prevent overfitting dropout regularization [51], with a dropout rate of 0.7 was used for each hidden layer. As loss function binary cross entropy was used.

5.8.3 Training

Due to an imbalance between the expected events of $t\bar{t}$ (96.77%) and tW (3.33%), class weights were used. At first all events ($t\bar{t}$ and tW) were shuffled and batched with a batchsize 1024 to ensure that some events per batch belonged to tW . This did not yield good results. To increase the amount of tW events per batch, the batches were created by selecting 512 events from each process. This approach has the disadvantage of a slower training. In order to make the training process faster, an imbalance of 1:3 was chosen, where a batch contains 512 tW events and 3·512 $t\bar{t}$ events. This was taken into account by setting the event weights accordingly.

5.8.4 Results

The output distribution of the classifier NN_{out} can be seen in Figure 26a. The tW events were assigned the class label 0 and the $t\bar{t}$ events were assigned the class label 1. In Figure 26b the shape plot of the NN_{out} can be seen, the shift towards 1 resides from the imbalance in the dataset.



(a) Output of the NN classifier, where the tW events were assigned class 0 and the $t\bar{t}$ events were assigned class 1.

(b) Shapes of the NN classifier output for the tW and $t\bar{t}$. The NN_{out} of tW is shifted towards 1 because of the imbalance between tW and $t\bar{t}$ events during training.

For the selection of events a cut on the NN_{out} of 0.39 was chosen, the number of data events after the cut is 5198 and the discrepancy is 21.3%. The number of events after the cut is shown in Table 8.

Table 8: Number of events after a cut of 0.39 on NN_{out} .

Process	Events
data	5198
$t\bar{t}$	4092
tW	927
fakes	35
$t\bar{t}V$	19

Figure 27 shows the distributions for $p_T(e)$, $p_T(e, b1)$, $|\eta(e)|$, and $|\eta(b1)|$ after cutting on the NN_{out} distribution. The plots for $p_T(e)$ show that the ratio of tW to $t\bar{t}$ events is better for high p_T , but the absolute number of tW events is higher in the low p_T region. The shape of the two variables is similar to the ones before the cut. For the two $|\eta|$ plots the shape changed from the plots without the cut, the NN classifier remove proportionally more events in the lower region. The shape for all four plots are similar to the ones in Figure 21, except for $p_T(b1)$, the classifier has more events in the high region for $t\bar{t}$ and tW . Overall, the four distributions look more similar to the ones after the cut by mini-max $m(b1)$ than after the cut by the AE.

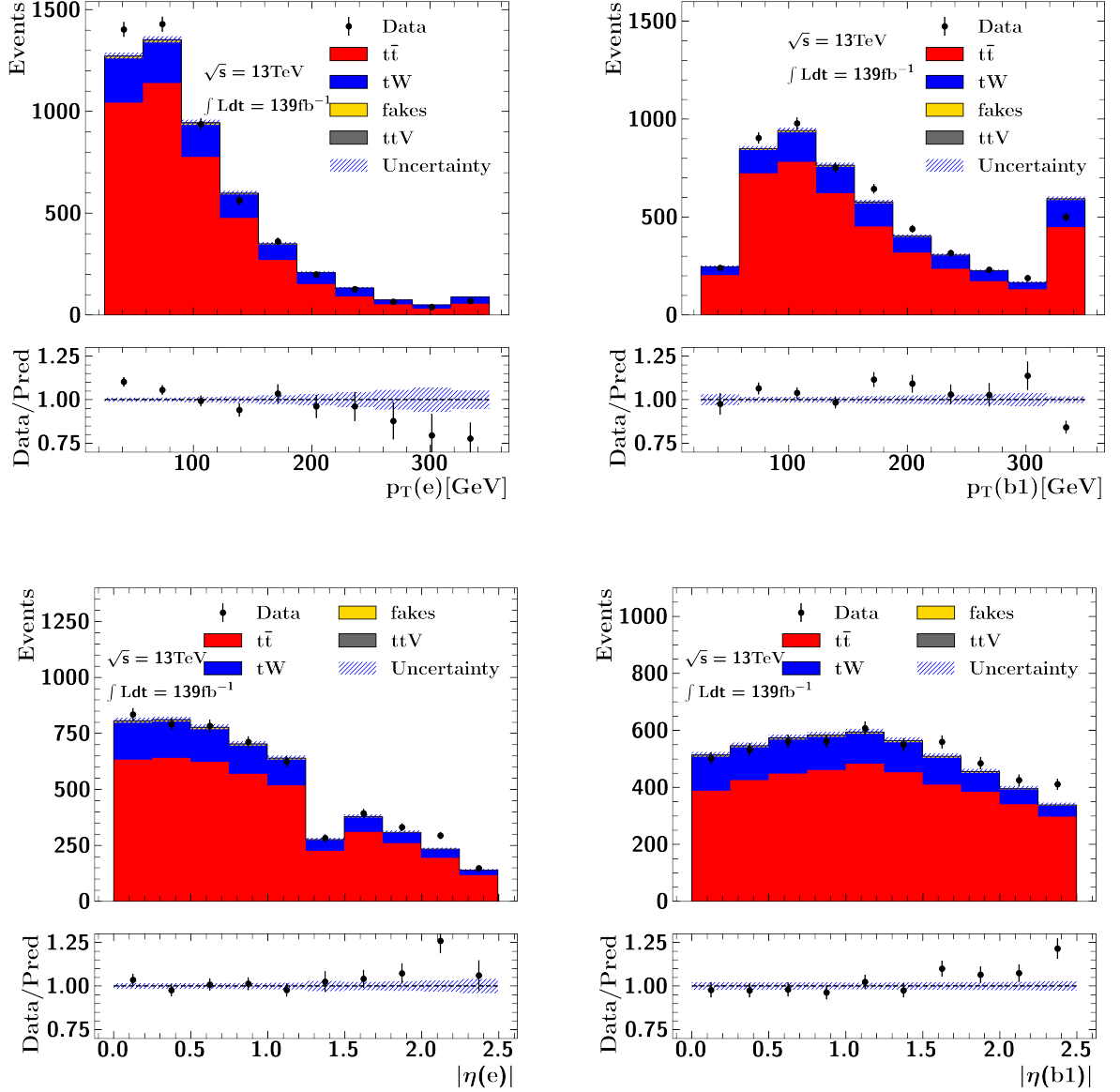


Figure 27: Plots of $p_T(e, b1)$ and $|\eta(e, b1)|$ after the cut with the classifier.

Figure 28 shows the distribution of E_T^{miss} , H_T , $m(e, b1)$, and $m(e, b2)$ after the cut with the classifier. Compared to the plots of the same variables for mini-max $m(b1)$ (Figure 22) and the AE (Figure 25), the shape of E_T^{miss} and H_T is closer to the shape of the variables after the mini-max $m(b1)$ cut. For the two masses the height of the second bump (originating from the off-shell produced Wb pair) is higher than for the AE, but not as pronounced as for mini-max $m(b1)$. This leads to the conclusion that the classifier internally has a “stronger” representation of mini-max $m(b1)$ than the AE.

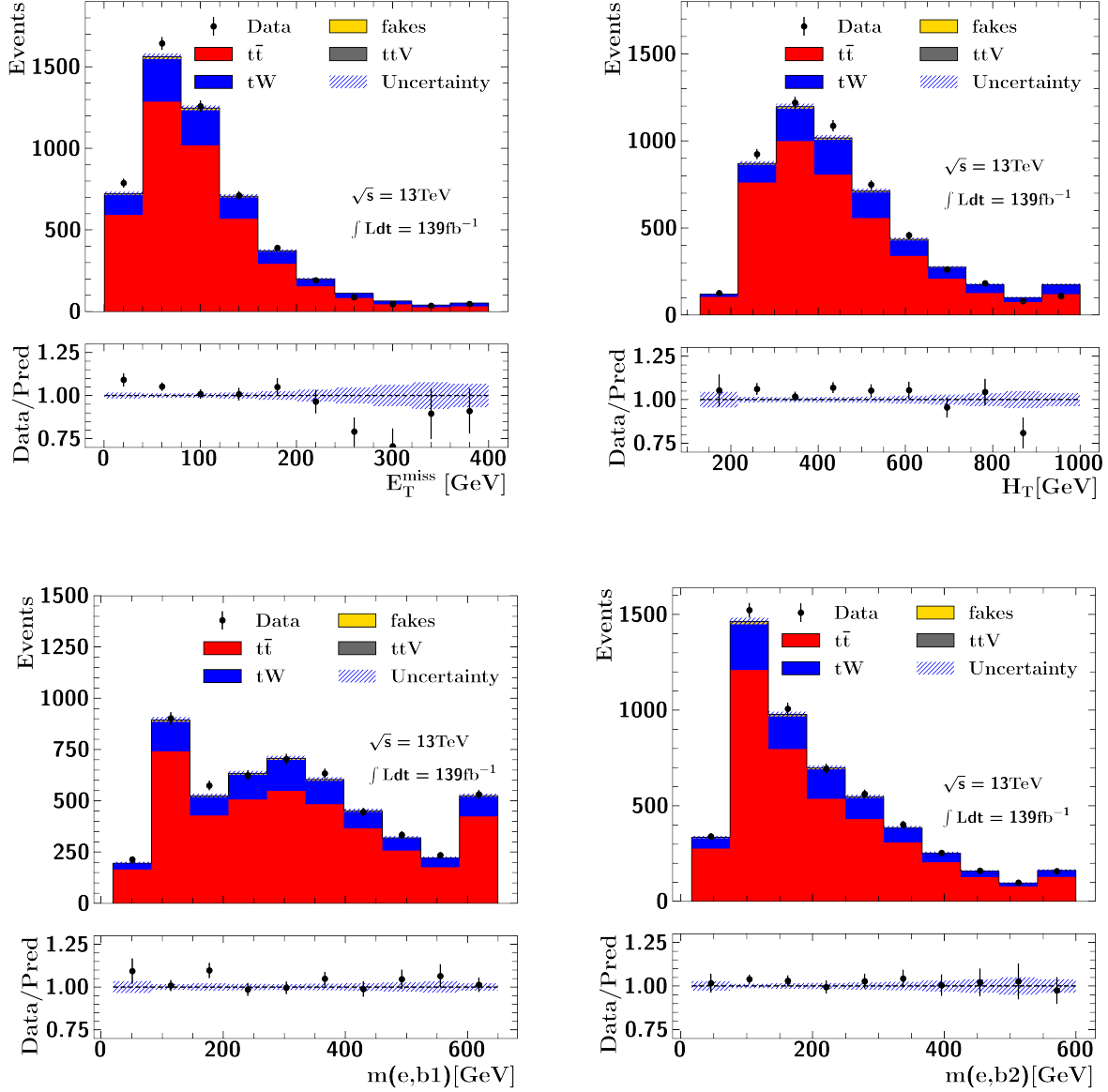


Figure 28: Plots of E_T^{miss} , H_T , $m(e, b1)$ and $m(e, b2)$ after the cut with the classifier.

5.8.5 Conclusion

In general, it can be expected that a system that was trained in a supervised manner outperforms a system that was trained in a semi-supervised manner. Hence it is expected that the classifier performs better than the AE. The AE increased the discrepancy from 5.4% to 16.5%, which is 4.8% less than the 21.3% achieved by the classifier. The increase comes mainly from a better representation of mini-max $m(bl)$ in the classifier NN, which can be seen above. The performance of the classifier is 0.6% lower than the one for mini-max $m(bl)$.

Overall, the classifier performs better than the AE but slightly worse than mini-max

$m(bl)$. But compared to the AE the classifier needs full information from $t\bar{t}$ and tW to be trained successfully which is not the case for the AE.

5.9 Comparison

In this section the results of the AE and mini-max $m(bl)$ are compared. No comparison between the results obtained with the NN classifier and any of the other algorithms is done. To get a better understanding of the difference in the selection of events between mini-max $m(bl)$ and the AE, three new selections are made. The selections are based on the AE and mini-max $m(bl)$. A selection by AE or mini-max $m(bl)$ means, selecting all events that are still left after applying the algorithm as described above (Table 3 and 7 respectively).

For the first selection mini-max $m(bl)$ and the AE are applied consecutively to the data, leaving only the events that are selected by both algorithms. The second selection takes all events that are selected by the AE, but removes those, that are also selected by the mini-max $m(bl)$ analysis. The third selection is the same as the second one only swapping mini-max $m(bl)$ and the AE. Figure 29 shows a graphical representation of the selections.

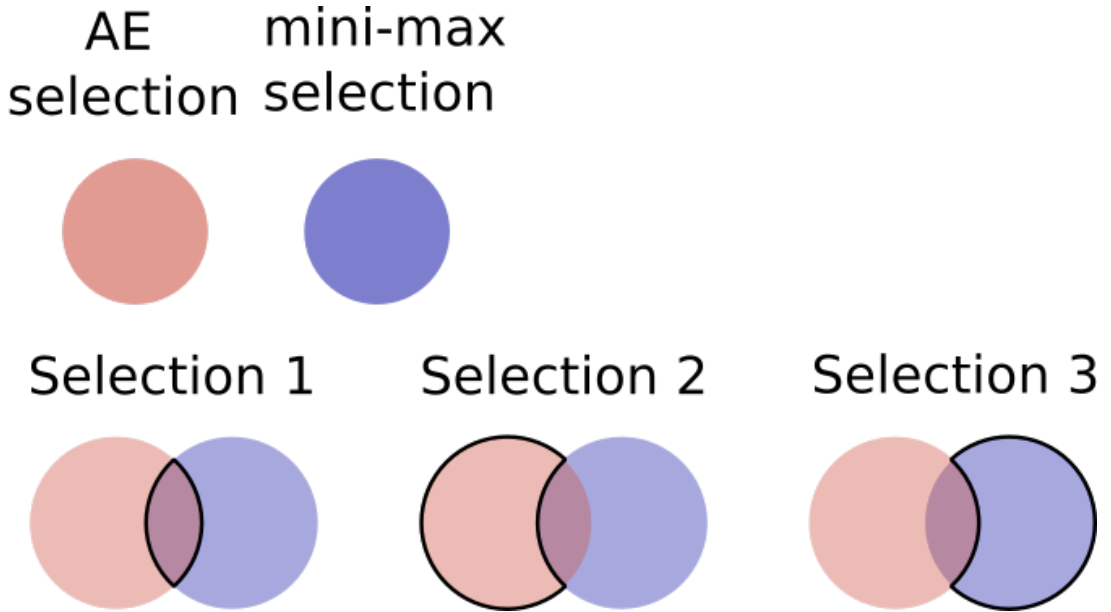


Figure 29: A graphical representation of the three selections made to investigate the difference between the event selection of the AE and mini-max $m(bl)$. The red circle represents the events selected by the AE and the blue circle represents the events selected by mini-max $m(bl)$. The black line represents the events used for each selection.

Selection 1 represents the intersection of the two algorithms. There are two possibilities for the properties of the selection. The first one is, that the selection does not resemble any

of the previous selections (here the AE selection or the mini-max $m(bl)$ selection). Thus, it can be concluded that the algorithms use very different information from the variables to select events. The second option is that the properties of the selection resemble one of the previous selections. In that case the resembled selection is a subset of the other one. Thus, it can be concluded that the algorithm of the resembled selection is contained in the other algorithm. These two possibilities open up a spectrum where a selection fits more to one or the other.

Selection 2 and 3 represent the exclusive set of events selected by each algorithm. The selection can be interpreted as the region that is unique to the algorithm.

Table 9 shows the number of event for each selection and process. Selection 1 retains roughly 45% of the events from the previous selections (the AE selection and the mini-max $m(lb)$ selection), leaving about 55% of the events for the exclusive selections.

Table 9: Number of events after each selection for each process. The number in brackets represent the number of events that are left after only the the AE or mini-max $m(bl)$, see Table 7 and Table 3 respectively.

Process	Events		
	Selection 1	Selection 2 (AE only)	Selection 3 (mini-max $m(bl)$ only)
data	2404	2982 (5386)	2816 (5220)
$t\bar{t}$	1799	2697 (4496)	2273 (4072)
tW	366	171 (537)	387 (753)
fakes	15	21 (36)	18 (33)
$t\bar{t}V$	9	9 (18)	10 (19)

From the information in Table 9 the discrepancies for each selection can be calculated. The discrepancies are shown in Table 10. The discrepancy of selection 1 is 25.2% which is an increase from the previous selections of the AE and mini-max $m(bl)$. As can be seen in Figure 20 and Figure 23, for mini-max $m(bl)$ and the AE respectively. he course of the discrepancy is nearly constant, it follows, that the algorithms partially select different events.

Table 10: Discrepancy for each selection.

Selection	Discrepancy
1	25.2%
2	9.6% (AE only 16.5%)
3	19.3% (mini-max $m(bl)$ only 21.9 %)

Because selection 1 and selection 2 combined give the events selected by the AE and the same goes for the combination of selection 1 and selection 3 for the events selected by mini-max, one can calculate the fraction of discrepancy that is unique to that algorithm. The fraction of the discrepancy, called fractional discrepancy is given by

$$D_f(s_{primary}, s_{subsequent}) = D_{s_{subsequent}} \cdot F(s_{primary}, s_{subsequent}), \quad (19)$$

where $s_{primary}$ is the selection done before $s_{subsequent}$ and $s_{subsequent}$ is the selection to get the fraction of $s_{primary}$ for and $F(s_{primary}, s_{subsequent})$ is given by $\frac{Events_{s_{subsequent}}}{Events_{s_{primary}}}$. Table 11 shows the fractional discrepancies. The values of $D_f(\text{AE, selection 1})$ and $D_f(\text{AE, selection 2})$ are 11.2% and 5.3% respectively.

$$D_f(\text{AE, selection 1}) = \left[1 - \frac{1799}{2404}\right] \cdot \frac{2404}{5386} = 11.2\% \quad (20)$$

$$D_f(\text{AE, selection 2}) = \left[1 - \frac{2697}{2982}\right] \cdot \frac{2982}{5386} = 5.3\% \quad (21)$$

Both fractional discrepancies added together give the total discrepancy of 16.5% obtained for the AE selection. Thus it follows, that about two thirds of the discrepancy of the AE is also selected by mini-max m(bl). The other third is unique to the AE and is not selected by mini-max m(bl). For mini-max m(bl) the fractional discrepancy for selection 1 is 11.6% and for selection 3 10.4%, so about 47.0% of the discrepancy from mini-max m(bl) is not covered by the AE.

$$D_f(\text{mini-max m(bl), selection 1}) = \left[1 - \frac{1799}{2404}\right] \cdot \frac{2404}{5220} = 11.6\% \quad (22)$$

$$D_f(\text{mini-max m(bl), selection 3}) = \left[1 - \frac{2273}{2816}\right] \cdot \frac{2816}{5220} = 10.4\% \quad (23)$$

Thus it follows, that about one third of the events selected by the AE are “mini-max” events, but the representation of mini-max m(bl) in the AE only covers 53.0% of mini-max m(bl).

Table 11: Fractional discrepancies for each selection.

$D_f(\text{AE, selection 1})$	11.2%
$D_f(\text{AE, selection 2})$	5.3%
$D_f(\text{mini-max, selection 1})$	11.6%
$D_f(\text{mini-max, selection 3})$	10.4%

To get a better understanding for the differences between the three selections and the physics covered by the algorithms specific kinematic variables are inspected next. Figure 30 show the distribution of $p_T(e)$ for the three selections. When looking at the higher region of most of the tW events are described by selection 1, where only selection 3 has some tW events left above 200 GeV and selection 2 does not have any tW events above 200 GeV. In the lower region both selection 2 and 3 have tW events. Because selection 2 only has tW events in the lower region, selection 3 has only has some tW events in the higher region and most of the tW events in the higher region are found in selection 1, it can be assumed, that both algorithms have the same sensitivity to high p_T events but differ in the event selection at low p_T .

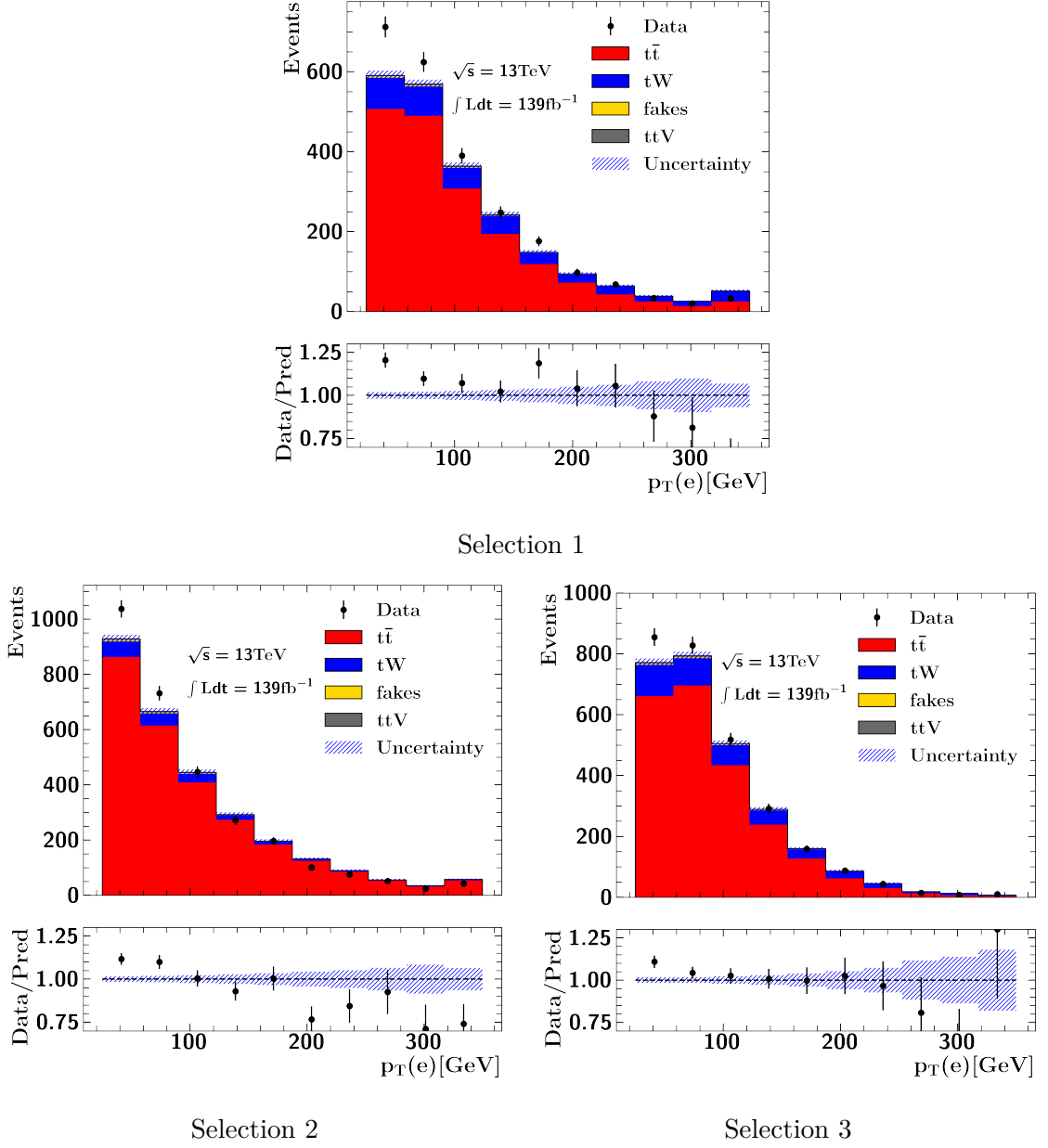


Figure 30: $p_T(e)$ for all three selections. The number of tW events for high p_T diminishes for selection 2 and 3, which means mini-max m(bl) and the AE select the same events for high p_T . For lower p_T both selections have events left, which means the two algorithms select different events at low p_T .

Figure 31 shows the distribution of $|\eta(b1)|$ for the three selections. Both selection 2 and 3 have tW events over the whole range, but the shape differs between the two. Selection 2 removed more events with a low $|\eta|$ than selection 3, but also has a lower sensitivity in that region.

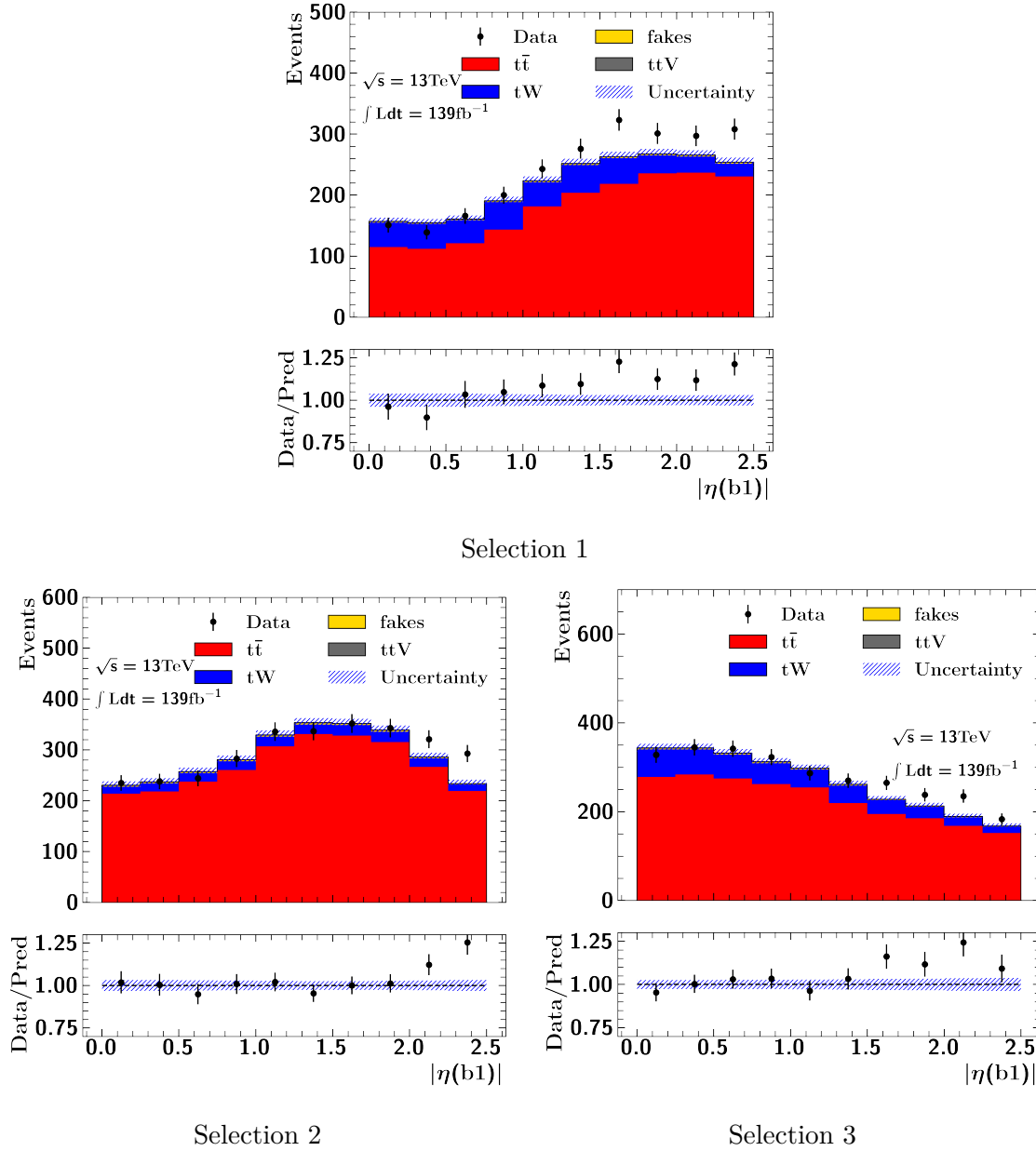


Figure 31: $|\eta|(e)$ for all three selections. The number of tW events for high $|\eta|$ diminishes for selection 2 and 3 which means mini-max $m(b1)$ and the AE select the same events for high $|\eta|$. For lower $|\eta|$ selection 3 has more events left than selection 2, which means that the mini-max $m(b1)$ selection is unique in the low $|\eta|$ region.

Figure 32 shows the distribution of $m(e, b1)$ for the three selections. Selection 1 looks similar in shape to $m(e, b1)$ after the cut by mini-max $m(b1)$ (Figure 22) with an enhanced sensitivity for high masses. Selection 2 and 3 have less tW events in the high region than after the cut of the respective algorithm. Selection 3 keeps the general shape of the mini-max $m(b1)$ cut with the two peaks, for the on-shell and off-shell b, l -pair. In the high region, specifically the last two bins are much smaller. For selection 2 it is the same, most

of the events that were removed come from the high mass region. The peak from the on-shell b,l -pair increased strongly, which means, that the AE has a higher sensitivity for low masses and mini-max $m(bl)$ a higher sensitivity for higher masses. Because the shape of selection 1 looks very similar to the selection of mini-max $m(bl)$ it can be concluded that the AE has an internal representation of mini-max $m(bl)$ but a “weaker” scale. This is confirmed by the plot of selection 3 where the shape of $m(e, b1)$ is preserved after removing the intersection with the AE.

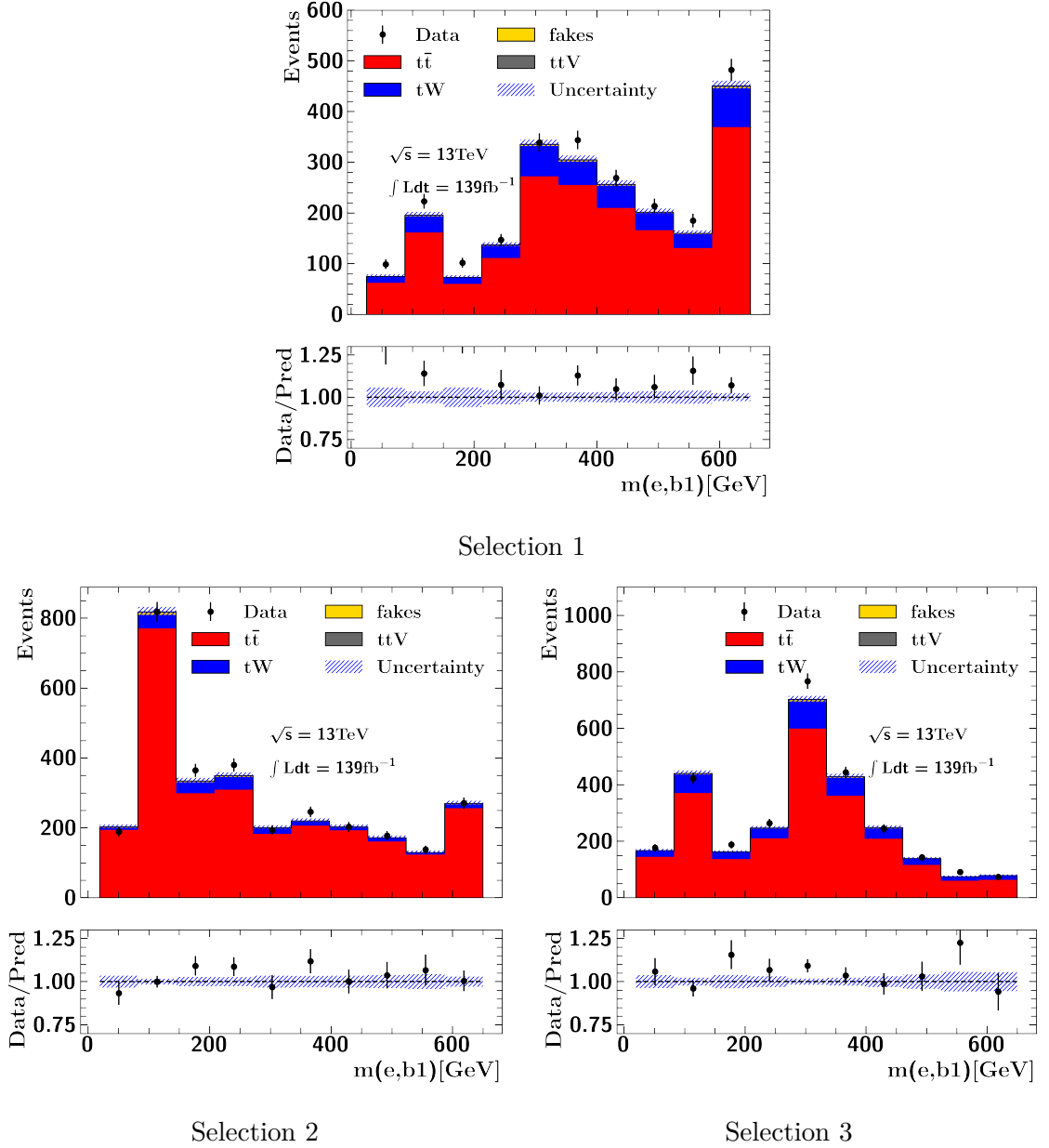


Figure 32: Shown is $m(e, b1)$ for all three selections. The shape of selection 3 is similar to the shape of mini-max $m(bl)$ and Selection 2 is missing the second peak, which comes from the off-shell b,l -pairs.

5.10 Conclusion

In this section three different models were tested at separating $t\bar{t}$ events from tW events. The metric used to evaluate the algorithms is the discrepancy after the corresponding cut. The discrepancy is equal to the fraction of events in data, that is not described by the $t\bar{t}$ simulation. To compare the algorithms at the same signal yield the number of data events after the cut had to be close to 5300 after the cut.

The first algorithm uses a variable called mini-max $m(b\bar{l})$, which is based on the physics for producing $t\bar{t}$ pairs and single top tW . It uses the fact, that the $b\bar{l}$ -pair from the tW process is produced off-shell from the top mass, which is not the case for the $t\bar{t}$ process. The shape of $m(l, b)$ is quite unique after the cut of mini-max $m(b\bar{l})$ with two peaks, one for the on-shell process and one for the off-shell process. The mini-max $m(b\bar{l})$ variable achieves a discrepancy of 21.9% after a cut with 5220 data events left.

The AE was trained semi-supervised, only using $t\bar{t}$ events, tW events were only used at testing time to evaluate the performance of the AE. The AE achieved a discrepancy of 16.5% with 5386 data events left after the cut.

To compare the semi-supervised AE against a fully supervised algorithm a binary classifier was trained. The classifier was trained on the $t\bar{t}$ and tW sample, it achieved a discrepancy of 21.3% with 5198 data events left after the cut.

By comparing the distributions of different variables after the cut of the respective algorithm it could be seen that the classifier had a “stronger” representation of mini-max $m(b\bar{l})$ than the AE. This was especially apparent in the distributions of $m(l, b)$, where the second peak coming from the off-shell $b\bar{l}$ -pair was more pronounced for the classifier than for the AE.

To get a better understanding of the AE and a better comparison to mini-max $m(b\bar{l})$ three selections were derived from the two algorithms. The first selection was the intersection of the selected events, the second and third were the exclusive selection of the AE and mini-max $m(b\bar{l})$ respectively. The first selection had a discrepancy of 25.2% with 2404 data events, the second selection had a discrepancy of 9.6% with 2982 data events, and the third selection had a discrepancy of 19.3% with 2816 data events. From the increased discrepancy for selection 1 it follows, that in the intersection the algorithms enhance each other and remove some of the $t\bar{t}$ events, that the other algorithm does not remove in their selection. Such a behavior could not come from the reduced number of events due to an automatic scaling of the algorithms, because the discrepancy curves for the higher regions of the respective variable ($m(\text{mini-max } m(b\bar{l}))$ and the reconstruction error) are constant, as can be seen in Figure 20 and Figure 23. To understand how much of the discrepancy of one algorithm is covered by the other one the fractional discrepancy D_f was calculated. The fractional discrepancies for the AE are $D_f(\text{selection 1, AE})=11.2\%$ and $D_f(\text{selection 2, AE})=11.2\%$.

AE)=5.3%. From this it follows, that two thirds of the discrepancy of the AE is covered by mini-max $m(\text{bl})$ and one third is unique to the AE. For mini-max $m(\text{bl})$ the fractional discrepancies are $D_f(\text{selection1}, \text{mini-max } m(\text{bl}))=11.6\%$ and $D_f(\text{selection3}, \text{mini-max } m(\text{bl}))=10.4\%$. That means about 55% of the discrepancy of mini-max $m(\text{bl})$ is covered by the AE and 45% are unique to mini-max $m(\text{bl})$. The events unique to the AE are in the lower region of the $m(l, b)$ variables and the events unique to mini-max $m(\text{bl})$ can be found in the low and mid region of the $m(l, b)$ variables.

In conclusion, the AE learned a “weaker” representation of mini-max $m(\text{bl})$, covering about 55% of mini-max $m(\text{bl})$, but it learned also some additional properties of the data, that are not included in mini-max $m(\text{bl})$. The discrepancy achieved by the AE is lower as than the discrepancy of mini-max $m(\text{bl})$, but the AE only used information of the known $t\bar{t}$ process and no information from the tW process.

6 Application of Autoencoders for $t\bar{t}$ +jets

AEs cannot only be used to search for a new signal, but also to test if a discrepancy between data and MC originates from a scaling error or a shape miss match. This section shows the application of an AE for such a problem. This section looks at a control region defined in [53], where most of the events come from $t\bar{t}$ +jets. The Feynman diagram at LO of $t\bar{t}$ +jets production can be seen in Figure 33.

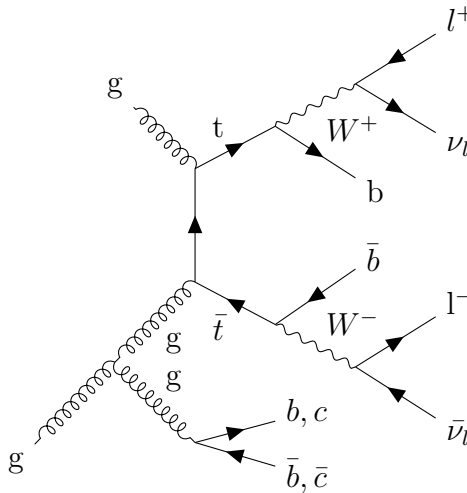


Figure 33: Feynman diagram for $t\bar{t}$ +jets at LO in the dilepton channel.

6.1 Data and simulated samples

This section is based on the information given in Ref. [53].

The data sample is the same as described in Section 5.2.

6.1.1 Monte Carlo samples

The following paragraphs present for the generated MC samples for each process. Table 12 gives an overview of the samples.

$t\bar{t}$ pairs

The $t\bar{t}$ sample is the same as described in Section 5.2. With the exception, that the $t\bar{t}$ +jet production is only simulated for $t\bar{t}$ +1 jet at NLO.

Single top

The single top events are modeled with POWHEGBOX [33, 34, 35] v2 generator, that provides MEs at NLO using the NNPDF3.0nlo [36] PDF. The functional form of the renormalization and factorization scale are set to the default scale. To handle the interference

with $t\bar{t}$ -pair production, the diagram removal scheme [43] was employed. Top quark decays are simulated at LO using `MADSPIN` [38, 39]. The events are interfaced with `PYTHIA8` (v8.230) [40], using the `A14` tune [41] and the `NNPDF3.0nnlo` [36] PDF set. Bottom and charm hadron decays are simulated using `EVTGEN` v1.6.0 [42].

Z+jets/Diboson

The Z+jets events were simulated with the parton shower MC generator `SHERPA` v2.2 [54]. As PDF the `NNPDF3.0nnlo` [36] set and a dedicated set of tuned parton-shower parameters developed by the `SHERPA` authors has been used.

Table 12: Overview of the simulated samples.

Process	ME Generator	ME PDF	Parton shower/ hadronization
$t\bar{t}$	<code>POWHEGBOX</code>	<code>NNPDF3.0nnlo</code>	<code>PYTHIA8</code>
single top	<code>POWHEGBOX</code>	<code>NNPDF3.0nnlo</code>	<code>PYTHIA8</code>
Z+jets	<code>SHERPA</code>	<code>NNPDF3.0nnlo</code>	<code>SHERPA</code>
diboson	<code>SHERPA</code>	<code>NNPDF3.0nnlo</code>	<code>SHERPA</code>

6.1.2 Object definition

This section is based on the information given in Ref. [53].

Electrons

Electron candidates have to belong to the quality group `TightLH` with a $p_T > 24$ GeV and a $|\eta_{cl}| < 2.47$. The region of $1.37 < |\eta_{cl}| < 1.52$ is excluded. The transverse impact parameter significance must be $|d_0/\sigma d_0| < 5$ and the longitudinal impact parameter $|\Delta z_0 \sin(\theta)| < 0.5$ mm. To reject candidates not coming from W bosons an isolation criteria is required. This is achieved by requiring $\Delta R = 0.2$ around the electron and the sum of the transverse momentum in the cone has to be of size 0.3 GeV. The parameters were chosen such, that the simulated isolation efficiency is 90%.

Muons

Muons are identified by tracking information from the muon detectors and the ID. Muon candidates have to belong to the quality group `TightLH` with a $p_T > 26$ GeV and

a $|\eta_{cl}| < 2.5$. Selected muons are required to fulfill additional quality cuts in the various tracking sub-detectors and must be classified as tight and isolated. The transverse impact parameter significance must be $|d_0/\sigma d_0| < 3$ and the longitudinal impact parameter $|\Delta z_0 \sin(\theta)| < 0.5$ mm. The isolation cuts were set to achieve a uniform isolation efficiency across p_T and η . Settings are the same as used for electrons (see above).

Jets

Jet candidates are reconstructed from topological clusters [45], with an anti- k_t algorithm [46] ($R=0.4$). All jets are required to have a $p_T > 25$ GeV and $|\eta| < 2.5$. To suppress jets with a large energy fraction from pileup collision vertices, the jet-vertex-tagger (JVT) [47] is used. The candidates are required to have a JVT value above 0.59 with a $p_T < 60$ GeV and $|\eta| < 2.4$. Jets that overlap with electron candidates within a cone of $\Delta R = 0.2$ are removed. If there is still an overlap of a jet with an electron candidate within $\Delta R = 0.4$ and the jet has a $p_T > 25$ GeV the electron candidate is removed. Jets that overlap with muons are removed within a radius of $\Delta R = 0.4$ if they have less than three tracks. If the muon satisfies $\Delta R(\text{muon, jet}) < 0.04$ and $p_T(\text{muon}) = 10$ GeV the muon is rejected if the jet has at least three tracks that originate from the primary vertex. All objects need to have a $p_T > 25$ GeV.

Heavy-flavour jet tagging

The b-jet tagging is the same as described in Section 5.3.1, with the only difference being the b-jet efficiency of 70%.

Missing transverse momentum

The missing transverse momentum E_T^{miss} comes from neutrinos, that cannot be measured or particles that could not be reconstructed. All momentum contributions not belonging to any of the objects listed above are reconstructed as the E_T^{miss} “soft term”.

6.2 Event selection

Events are required to have at least three jets with a $p_T > 30$ GeV. Furthermore, three of the jets have to be b-tagged jets with a working point of 80%. Additionally, two leptons were required with one being an electron and one a muon, or both being an electron or a muon. All muons need to have a $p_T > 25$ GeV. E_T^{miss} had to be above 30 GeV. The number of events after the selection for all processes are listed in Table 13. The discrepancy between data and the MC datasets is 17.2%.

Table 13: Number of events for each dataset after the selection described above.

Process	Events
$t\bar{t}$	26412
Z+jets	2177
single top	805
diboson	113
sum	29507
data	35648

6.3 Kinematic properties

In this section the distributions of kinematic variables are looked at to get a better understanding of them. Plots for all 61 variables that were used as input for the AE were generated, but to keep this section concise only a selection of 6 plots is shown. All other plots can be found in the Appendix B.1. A list of the plotted variables is given in Table 14.

Figure 34 shows the distribution of p_T for the leading lepton (l1), leading jet (j1) and the leading b-jet (b1). The distribution of $p_T(j1)$ is slightly more shifted towards higher values. From this it follows that not all primary jets are also b-jets, but due to the shift only being small most of the primary jets are b-jets. All three plot show a similar shape for the ratio between data and predicted events, with higher p_T the discrepancy diminishes.

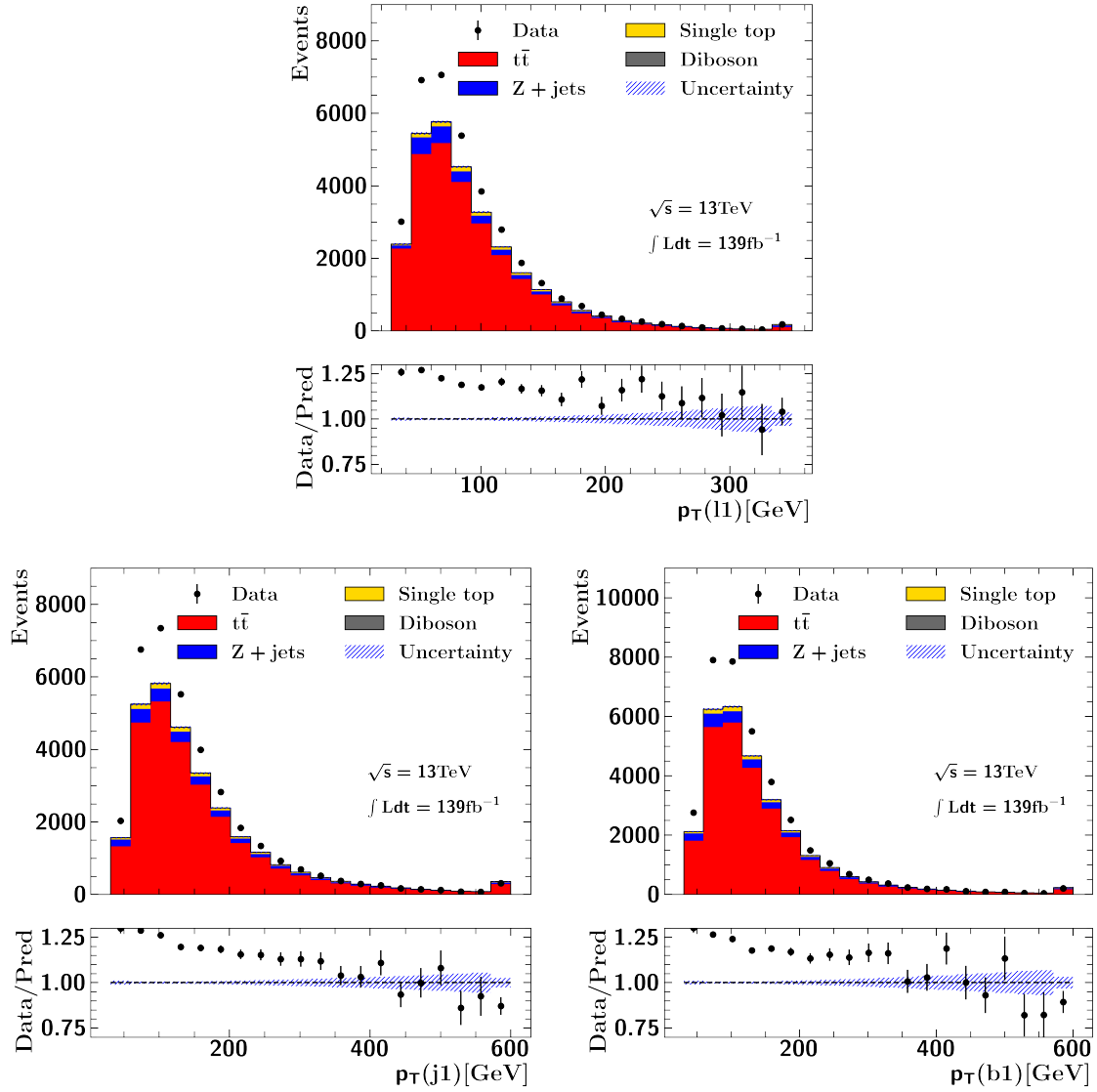


Figure 34: Distributions of p_T for l_1 , j_1 and b_1 . Due to the fact, that not all primary jets are also b-jets, the tail of $p_T(j_1)$ is longer.

Figure 35 shows the distribution of $|\eta|$ for the leading lepton, the leading jet and the leading b-jet. The distributions for the jet and b-jet have the same shape. The ratio between data and predicted events shows no trend as can be seen in Figure 34, the ratio is constant within some variation. For the leading lepton the value of data/predicted is also near constant except for the last two bins, where the value moves closer to 1.

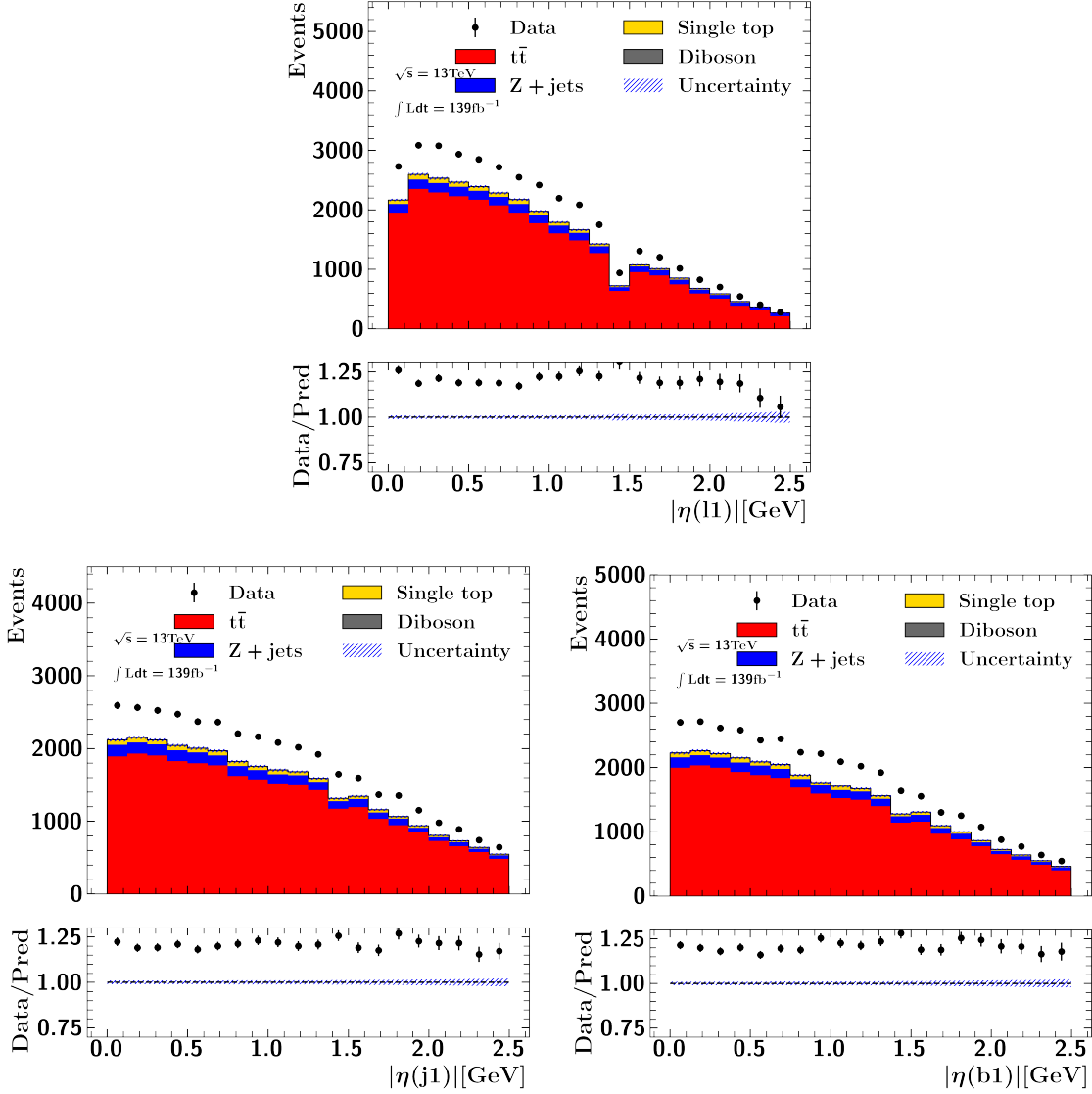


Figure 35: Distribution of $|\eta|$ for l1, b1 and j1. The distribution for j1 and b1 looks similar.

6.4 Autoencoder

As input for the AE, two datasets were generated. This was done to lower the number of inputs. The two datasets differ in the selection of the jets. Each dataset has 36 variables with 11 shared variables from the two leptons and E_T^{miss} . The first dataset uses the three primary jets and the second one uses the three b-jets. Table 14 shows the inputs for the two datasets.

Table 14: List of variables that were used as input for the AE. The first two rows are used in both datasets (indicated by “Both datasets”), the third and fourth row where only used in the jet dataset and the fifth and sixth row where only uses for the b-jet dataset. $H_{T,j,b}$ denote that the H_T belongs to the jet or b-jet selection respectively.

Both datasets			
$p_T(l_{1,2})$	$E(l_{1,2})$	$\eta(l_{1,2})$	$\phi(l_{1,2})$
$p_T(E_T^{miss})$	$E(E_T^{miss})$		$\phi(E_T^{miss})$
Jet dataset			
$p_T(j_{1,2,3})$	$E(j_{1,2,3})$	$\eta(j_{1,2,3})$	$\phi(j_{1,2,3})$
$M(l_{1,2}, j_{1,2,3})$	$\Delta R(l_{1,2}, j_{1,2,3})$	H_{T_j}	
B-Jet dataset			
$p_T(b_{1,2,3})$	$E(b_{1,2,3})$	$\eta(b_{1,2,3})$	$\phi(b_{1,2,3})$
$M(l_{1,2}, b_{1,2,3})$	$\Delta R(l_{1,2}, b_{1,2,3})$	H_{T_b}	

The AE uses a symmetric structure around the latent dimension. The input layer has a size of 36 nodes and the latent dimension has a dimension of 24. The compression in the latent dimension is $24/36 = 66.67\%$. The AE uses four layers for the encoder and decoder, the node reduction/increase per layer is 3, which results in the following structure $36 \rightarrow 33 \rightarrow 30 \rightarrow 27 \rightarrow 24 \rightarrow 27 \rightarrow 30 \rightarrow 33 \rightarrow 36$. This structure was found to produce the best results for both selections (jet and b-jet selection) after testing different number of layers and number of nodes reduced per layer. For the different hyperparameters of the AE a parameter study was done, Table 15 shows the hyperparameters and the values that were tested. The training took about 15 minutes with $2.4 \cdot 10^5$ events, to prevent over training early stopping was used for the validation loss with a patience of 15.

Table 15: Hyperparameter that were searched when optimizing the AE. Final set of parameters is given below in Table 16

Parameter	Values
Activation function	sigmoid, relu, leaky relu, selu
Learning rate	0.001, 0.0001, 0.00001
Optimizer	SGD, Adam
Loss function	MAE, MSE
Architecture	see above

Table 16 shows the final set of hyperparameters that produced the best results.

Table 16: Final set of hyperparameters for the AE, that produced the best results. Both the AE for the jet dataset and for the b-jet dataset have same set of hyperparameters.

Parameter	Values
Activation function	selu
Learning rate	0.0001
Optimizer	Adam
Loss function	MSE
Architecture	see above

6.5 Result and conclusion

The reconstruction error for the both datasets of the final AEs are shown in Figure 36. The plot on the left shows the reconstruction error for the b-jet dataset and the plot on the right for the jet dataset. The black line in the lower subplots indicates the baseline discrepancy of 17.2%. The values of the ratio between data and predicted events show a slight trend towards 1 with higher reconstruction error, only for the last bin the value is very close to 1. This trend can also be seen in Figure 34 and does not indicate any anomalies. Overall, the structure of the ratio between data and predicted events is the same one as the one in the plots of the kinematic variables. This suggests that there is no evidence for new signal in the data and that the discrepancy between data and MC originates from a shape miss match and/or a scaling error.

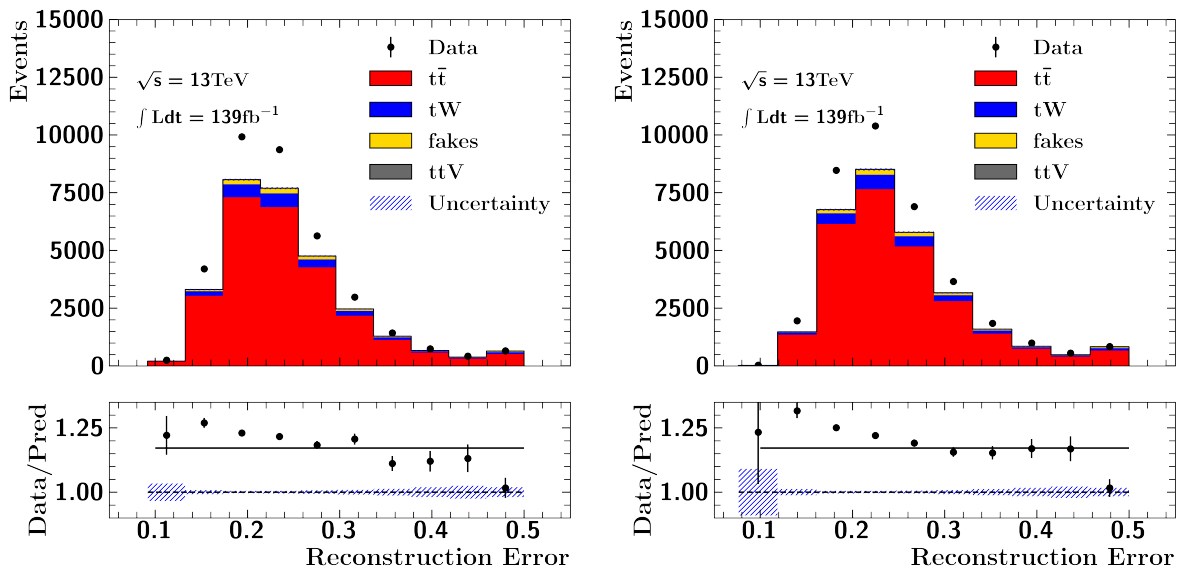


Figure 36: The reconstruction error for the b-jet dataset (left) and the primary jet dataset (right). Black line represents the base line discrepancy of 0.172

7 Conclusion and outlook

In this master thesis, the application of AEs for new physics signal searches was investigated. For the investigation, data from Run 2 of the LHC was used. The data was collected at a center-of-mass energy of $\sqrt{s}=13$ TeV with an integrated luminosity of 139 fb^{-1} .

In the first part, events from $t\bar{t}$ pair production were used. The selection of events required exactly two leptons in the final state with one being an electron and the other a muon with opposite charge. Furthermore, exactly two b-jets were required. After the selection 175460 data events and 165995 $t\bar{t}$ MC events were left. The goal was to find the tW events that make up the biggest part of the missing events between data and $t\bar{t}$ MC. tW was chosen as signal due to the purity of $t\bar{t}$ and tW in the investigated phase space. The metric used to evaluate the performance of an algorithm is the discrepancy, which is equal to the amount of data not described by the simulation of the $t\bar{t}$ process. The baseline discrepancy is 5.4%. For comparison, the physically motivated variable mini-max $m(\text{lb})$ was used, as well as a NN classifier to compare the performance of a semi supervised algorithm against a fully supervised algorithm. All algorithms were compared at same signal yield with close to 5300 data events. mini-max achieved a discrepancy of 21.9% with 5220 data events, the AE achieved a discrepancy of 16.5% with 5386 data events, and the NN classifier achieved a discrepancy of 21.3% with 5198 data events. The NN classifier learned a better representation of mini-max than the AE, which could be seen in the plots of the invariant masses after the cuts.

To improve the understanding of the differences between the AE and mini-max three selections were made (named selection 1-3). Selection 1 contains the intersection of the original selections done by AE and mini-max. Selection 2 and selection 3 are the exclusive events selected by the AE and mini-max respectively. Selection 1 has a discrepancy of 25.2% with 2404 data events, selection 2 9.6% with 2982 data events, and selection 3 19.3% with 2816 data events. It follows, that the two algorithms use different criteria in the intersection that select the same tW events, but remove $t\bar{t}$ events the other algorithm would select. To get the fraction of the discrepancy for selection 1-3 of the original selection (AE or mini-max), the fractional discrepancies were calculated. The fractional discrepancy for selection 1 and the AE selection is $D_f(\text{selection 1, AE})=11.2\%$ and $D_f(\text{selection 2, AE})=5.3\%$ for the AE and selection 2. From this follows, that about one third of the discrepancy of the AE is covered by mini-max and one third is unique to the AE. For mini-max the fractional discrepancies are $D_f(\text{selection 1, mini-max})=11.6\%$ and $D_f(\text{selection 3, mini-max})=10.4\%$. Thus, about 55% of the discrepancy of mini-max is covered by the AE and 45% are unique to mini-max. Furthermore, looking at the distributions of the invariant masses of the b,l-pairs for selection 1-3 confirms that the AE learned a partial representation of mini-max. The learned representation of mini-max covers about 55% of

mini-max and was learned only using information of the $t\bar{t}$ process.

In the second part of this work AEs were used to investigate the discrepancy in a region of $t\bar{t}$ that required exactly two leptons, at least 3 jets, and exactly 3 b-jets in the final state. For the datasets two selections of variables were made. One selecting the three primary jets and one selecting the three b-jets. For both selections AEs were trained. The plots of the reconstruction error suggests that no new signal is the reason for the discrepancy, but that the discrepancy originates from a scaling error and/or shape miss match.

To conclude, AEs are capable of finding new physics signatures only using known processes, without any theoretical bias. However, as it is expected that the separation between data and simulation lacks in performance compared to a fully supervised model. Moreover, the AE is limited in the representation of mini-max, where it only selects a subset of the events selected by mini-max. Despite these limitations, the AE selects an additional region, with an increased discrepancy, which is not covered by mini-max. A study of that region would be of interest for future work. Furthermore, AEs can be used to investigate if errors occurred during the generation of simulated samples. Although, AEs are capable of increasing the discrepancy between MC and data, as shown in this work, their primary use should be to find hints at the underlying physics of the discrepancy. From there, the information can be used to develop more powerful methods for separating signal and background. AEs are quite inexpensive from a computational standpoint with a training time of minutes to hours. With the possibility to perform hyperparameter searches in parallel, AEs can be used to get a first insight into a dataset with little time and computational cost.

References

- [1] D. J. Griffiths. *Introduction to elementary particles; 2nd rev. version*. Physics textbook. New York, NY: Wiley, 2008. URL: <https://cds.cern.ch/record/111880>.
- [2] M. E. Peskin and D. V. Schroeder. *An Introduction to quantum field theory*. Reading, USA: Addison-Wesley, 1995. ISBN: 9780201503975, 0201503972. URL: <http://www.slac.stanford.edu/~mpeskin/QFT.html>.
- [3] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Phys. Lett.* B716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].
- [4] S. Chatrchyan et al. “Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC”. In: *Phys. Lett.* B716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].
- [5] M. Tanabashi et al. “Review of Particle Physics”. In: *Phys. Rev. D* 98 (3 2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001. URL: <https://link.aps.org/doi/10.1103/PhysRevD.98.030001>.
- [6] Y. Ashie et al. “A Measurement of atmospheric neutrino oscillation parameters by SUPER-KAMIOKANDE I”. In: *Phys. Rev.* D71 (2005), p. 112005. DOI: 10.1103/PhysRevD.71.112005. arXiv: hep-ex/0501064 [hep-ex].
- [7] A. Buckley et al. “General-purpose event generators for LHC physics”. In: *Phys. Rept.* 504 (2011), pp. 145–233. DOI: 10.1016/j.physrep.2011.03.005. arXiv: 1101.2599 [hep-ph].
- [8] J. C. Collins, D. E. Soper, and G. F. Sterman. “Factorization of Hard Processes in QCD”. In: *Adv. Ser. Direct. High Energy Phys.* 5 (1989), pp. 1–91. DOI: 10.1142/9789814503266_0001. arXiv: hep-ph/0409313 [hep-ph].
- [9] R. K. Ellis, W. J. Stirling, and B. R. Webber. “QCD and collider physics”. In: *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.* 8 (1996), pp. 1–435.
- [10] R. D. Ball et al. “Parton distributions from high-precision collider data”. In: *Eur. Phys. J. C* 77.10 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph].
- [11] F. Abe et al. “Observation of top quark production in $\bar{p}p$ collisions”. In: *Phys. Rev. Lett.* 74 (1995), pp. 2626–2631. DOI: 10.1103/PhysRevLett.74.2626. arXiv: hep-ex/9503002 [hep-ex].

- [12] S. Abachi et al. “Observation of the top quark”. In: *Phys. Rev. Lett.* 74 (1995), pp. 2632–2637. DOI: 10.1103/PhysRevLett.74.2632. arXiv: hep-ex/9503003 [hep-ex].
- [13] M. Aaboud et al. “Search for pair production of vector-like top quarks in events with one lepton, jets, and missing transverse momentum in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector”. In: *JHEP* 08 (2017), p. 052. DOI: 10.1007/JHEP08(2017)052. arXiv: 1705.10751 [hep-ex].
- [14] L. Evans and P. Bryant. “LHC Machine”. In: *Journal of Instrumentation* 3.08 (2008), S08001–S08001. DOI: 10.1088/1748-0221/3/08/S08001. URL: <https://doi.org/10.1088/1748-0221/3/08/S08001>.
- [15] K. Aamodt et al. “The ALICE experiment at the CERN LHC”. In: *JINST* 3 (2008), S08002. DOI: 10.1088/1748-0221/3/08/S08002.
- [16] L. Evans and P. Bryant. “LHC Machine”. In: *JINST* 3 (2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001.
- [17] S. Chatrchyan et al. “The CMS Experiment at the CERN LHC”. In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.
- [18] A. A. Alves Jr. et al. “The LHCb Detector at the LHC”. In: *JINST* 3 (2008), S08005. DOI: 10.1088/1748-0221/3/08/S08005.
- [19] C. Lefèvre. “The CERN accelerator complex. Complexe des accélérateurs du CERN”. 2008. URL: <https://cds.cern.ch/record/1260465>.
- [20] A. Collaboration. *LuminosityPublicResultsRun2*. URL: <https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2>.
- [21] J. Pequeno. “Computer generated image of the whole ATLAS detector”. 2008. URL: <https://cds.cern.ch/record/1095924>.
- [22] M. Aaboud et al. “Performance of the ATLAS Trigger System in 2015”. In: *Eur. Phys. J. C* 77.5 (2017), p. 317. DOI: 10.1140/epjc/s10052-017-4852-3. arXiv: 1611.09661 [hep-ex].
- [23] J. Pequeno. “Event Cross Section in a computer generated image of the ATLAS detector.” 2008. URL: <https://cds.cern.ch/record/1096081>.
- [24] B. C. Csáji. “Approximation with Artificial Neural Networks”. In: 2001.
- [25] G. Aad et al. “Measurement of the Higgs boson mass from the $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^* \rightarrow 4\ell$ channels with the ATLAS detector using 25 fb^{-1} of pp collision data”. In: *Phys. Rev. D* 90.5 (2014), p. 052004. DOI: 10.1103/PhysRevD.90.052004. arXiv: 1406.3827 [hep-ex].

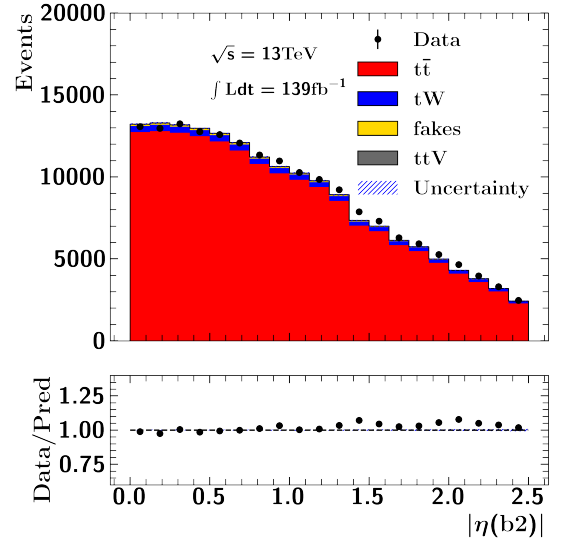
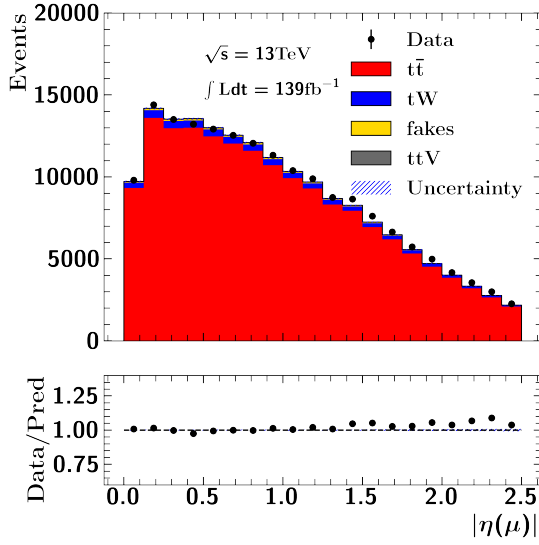
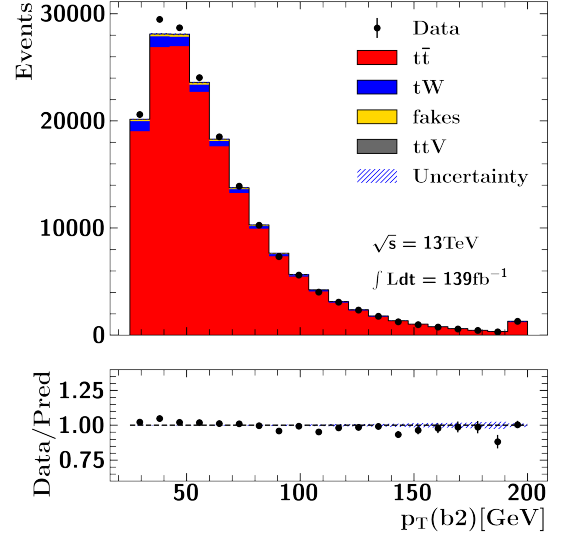
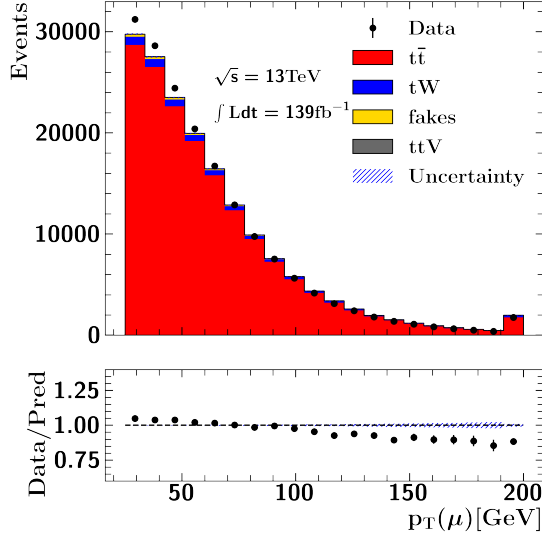
- [26] R. Benbrik et al. “Signatures of vector-like top partners decaying into new neutral scalar or pseudoscalar bosons”. In: (2019). arXiv: 1907.05929 [hep-ph].
- [27] G. Aad et al. “Search for flavour-changing neutral currents in processes with one top quark and a photon using 81 fb⁻¹ of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS experiment”. In: *Phys. Lett. B* 800 (2020), p. 135082. DOI: 10.1016/j.physletb.2019.135082. arXiv: 1908.08461 [hep-ex].
- [28] M. Aaboud et al. “A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment”. In: *Eur. Phys. J. C* 79.2 (2019), p. 120. DOI: 10.1140/epjc/s10052-019-6540-y. arXiv: 1807.07447 [hep-ex].
- [29] T. C. Herwig, T. Eifert, and B. P. Nachman. *Quantum Interference Between Single and Doubly Resonant Top Quark Production*. Tech. rep. ATL-COM-PHYS-2017-037. Geneva: CERN, 2017. URL: <https://cds.cern.ch/record/2243038>.
- [30] D. Hirschbuehl et al. *Measurements of observables sensitive to colour reconnection in $t\bar{t}$ dilepton events with the ATLAS detector at $\sqrt{s} = 13$ TeV*. Tech. rep. ATL-COM-PHYS-2019-888. Geneva: CERN, 2019. URL: <https://cds.cern.ch/record/2682602>.
- [31] L. Sartini et al. “Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment”. In: *Nuclear Instruments and Methods in Physics Research A* (2010, month =).
- [32] G. Aad et al. “The ATLAS Simulation Infrastructure”. In: *Eur. Phys. J. C* 70 (2010), pp. 823–874. DOI: 10.1140/epjc/s10052-010-1429-9. arXiv: 1005.4568 [physics.ins-det].
- [33] S. Frixione, P. Nason, and G. Ridolfi. “A Positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction”. In: *JHEP* 09 (2007), p. 126. DOI: 10.1088/1126-6708/2007/09/126. arXiv: 0707.3088 [hep-ph].
- [34] P. Nason. “A New method for combining NLO QCD with shower Monte Carlo algorithms”. In: *JHEP* 11 (2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040. arXiv: hep-ph/0409146 [hep-ph].
- [35] S. Frixione, P. Nason, and C. Oleari. “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method”. In: *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph].
- [36] R. D. Ball et al. “Parton distributions for the LHC Run II”. In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

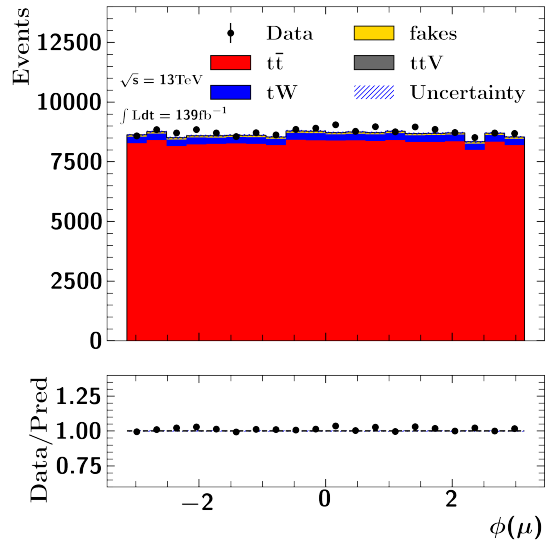
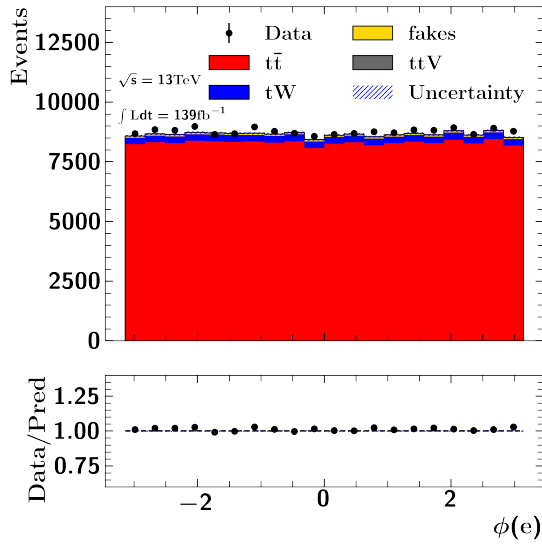
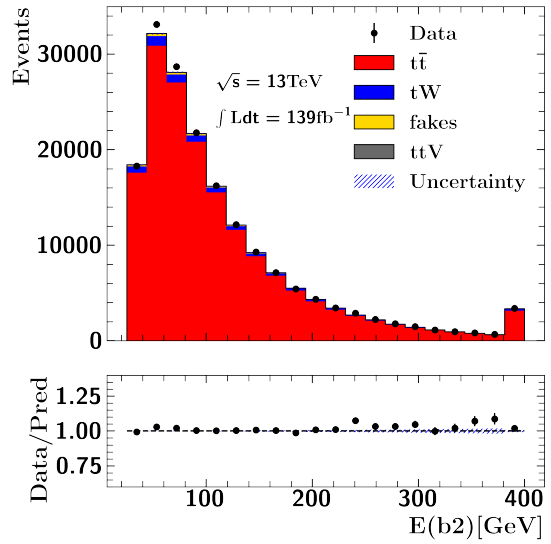
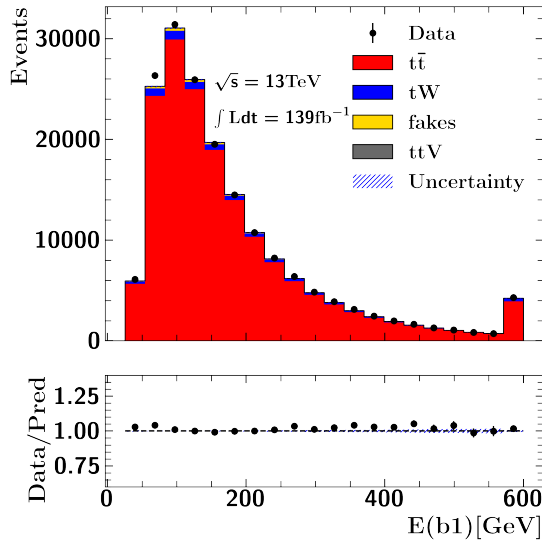
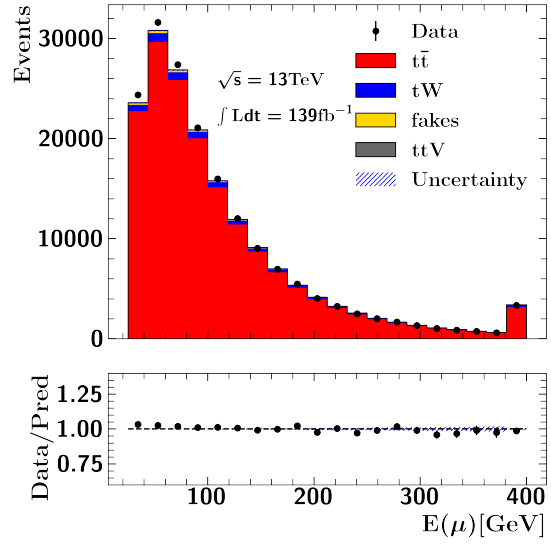
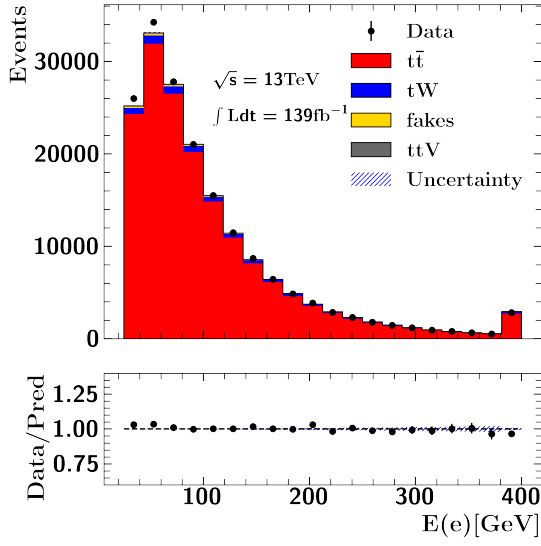
-
- [37] *Studies on top-quark Monte Carlo modelling for Top2016*. Tech. rep. ATL-PHYS-PUB-2016-020. Geneva: CERN, 2016. URL: <https://cds.cern.ch/record/2216168>.
- [38] S. Frixione et al. “Angular correlations of lepton pairs from vector boson and top quark decays in Monte Carlo simulations”. In: *Journal of High Energy Physics* 2007.04 (2007), pp. 081–081. DOI: 10.1088/1126-6708/2007/04/081. URL: <https://doi.org/10.1088%2F1126-6708%2F2007%2F04%2F081>.
- [39] P. Artoisenet et al. “Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations”. In: *JHEP* 03 (2013), p. 015. DOI: 10.1007/JHEP03(2013)015. arXiv: 1212.3460 [hep-ph].
- [40] T. Sjöstrand et al. “An Introduction to PYTHIA 8.2”. In: *Comput. Phys. Commun.* 191 (2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph].
- [41] *ATLAS Run 1 Pythia8 tunes*. Tech. rep. ATL-PHYS-PUB-2014-021. Geneva: CERN, 2014. URL: <https://cds.cern.ch/record/1966419>.
- [42] D. J. Lange. “The EvtGen particle decay simulation package”. In: *Nucl. Instrum. Meth.* A462 (2001), pp. 152–155. DOI: 10.1016/S0168-9002(01)00089-4.
- [43] S. Frixione et al. “Single-top hadroproduction in association with a W boson”. In: *JHEP* 07 (2008), p. 029. DOI: 10.1088/1126-6708/2008/07/029. arXiv: 0805.3067 [hep-ph].
- [44] J. Alwall et al. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”. In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].
- [45] G. Aad et al. “Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1”. In: *Eur. Phys. J.* C77 (2017), p. 490. DOI: 10.1140/epjc/s10052-017-5004-5. arXiv: 1603.02934 [hep-ex].
- [46] M. Cacciari, G. P. Salam, and G. Soyez. “The anti- k_t jet clustering algorithm”. In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].
- [47] *Constituent-level pile-up mitigation techniques in ATLAS*. Tech. rep. ATLAS-CONF-2017-065. Geneva: CERN, 2017. URL: <https://cds.cern.ch/record/2281055>.
- [48] *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*. Tech. rep. ATL-PHYS-PUB-2017-013. Geneva: CERN, 2017. URL: <https://cds.cern.ch/record/2273281>.

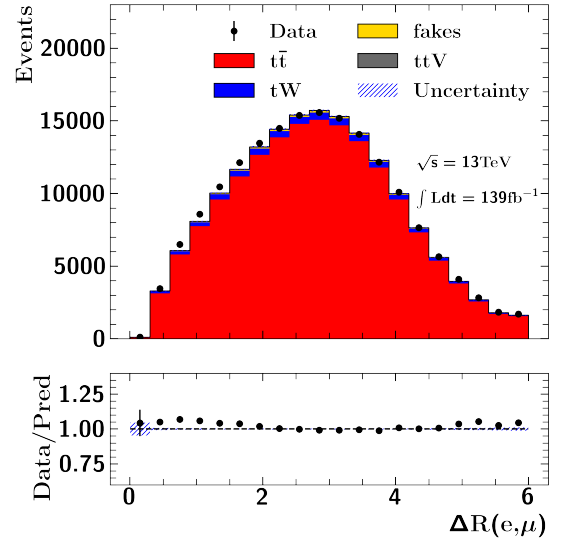
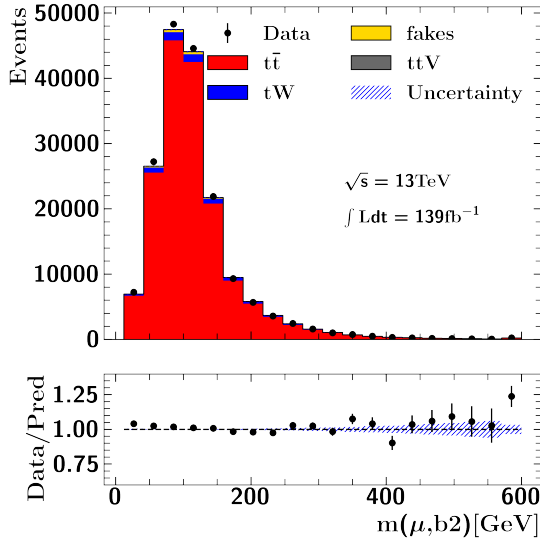
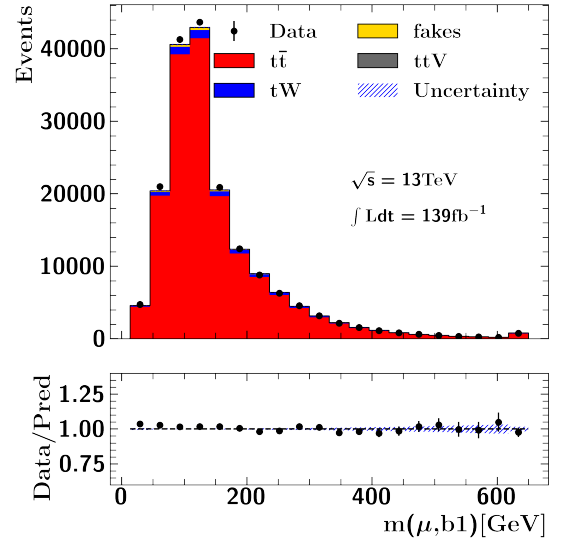
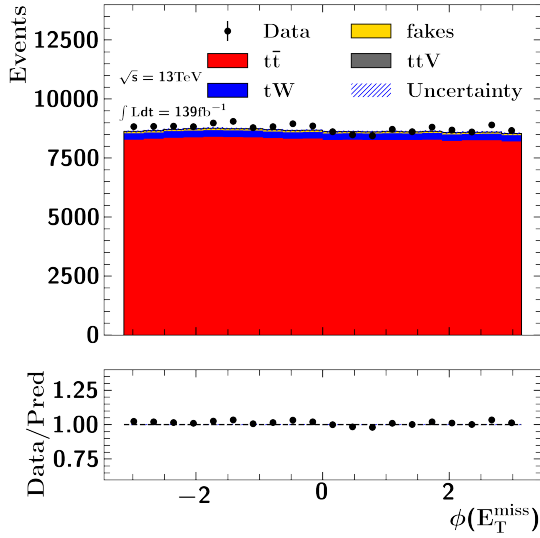
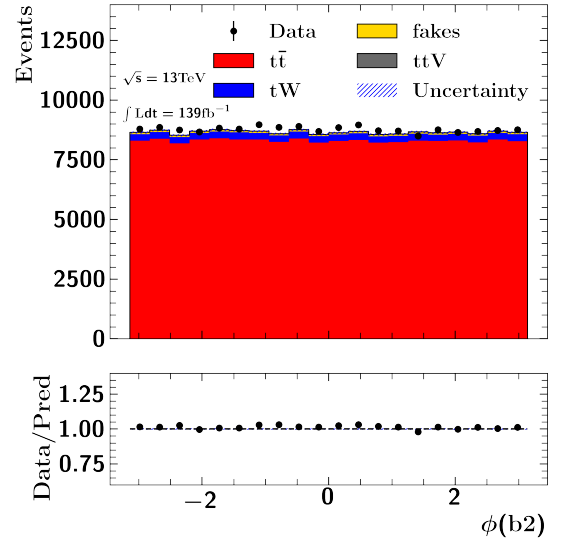
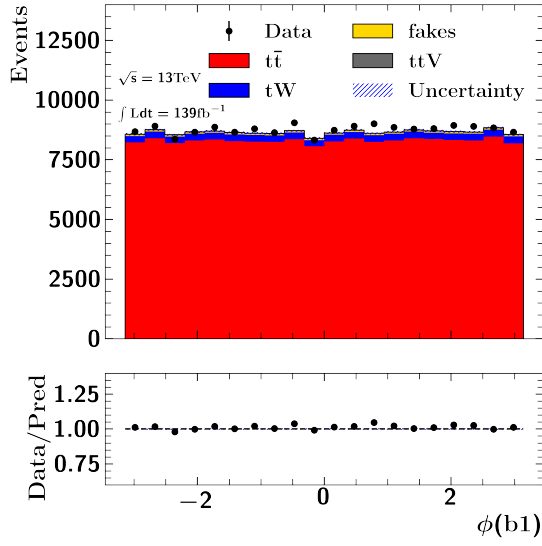
-
- [49] R. M. Sakia. “The Box-Cox Transformation Technique: A Review”. English. In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 41.2 (1992), pp. 169–178. ISSN: 00390526. URL: <http://www.jstor.org/stable/2348250>.
- [50] G. Klambauer et al. “Self-Normalizing Neural Networks”. In: *CoRR* abs/1706.02515 (2017). arXiv: 1706.02515. URL: <http://arxiv.org/abs/1706.02515>.
- [51] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016.
- [52] M. Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: <https://www.tensorflow.org/>.
- [53] G. Gilles et al. *Search for flavour-changing neutral current couplings between the top-quark and the Higgs boson in the $H \rightarrow b\bar{b}$ decay channel with the ATLAS detector at the LHC*. Tech. rep. ATL-COM-PHYS-2018-1500. Geneva: CERN, 2918. URL: <https://cds.cern.ch/record/2645330>.
- [54] T. Gleisberg et al. “Event generation with SHERPA 1.1”. In: *JHEP* 02 (2009), p. 007. DOI: 10.1088/1126-6708/2009/02/007. arXiv: 0811.4622 [hep-ph].

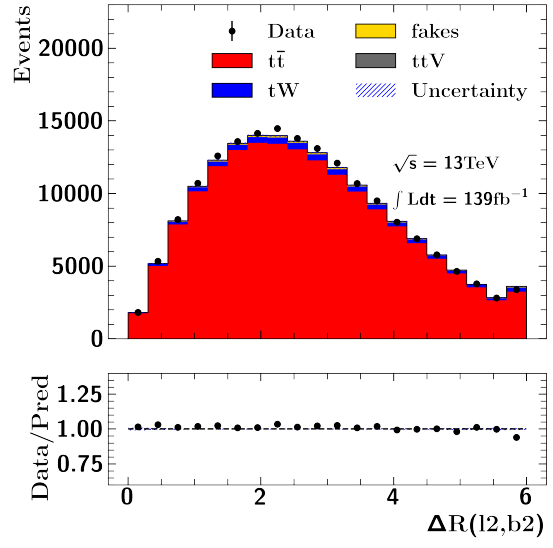
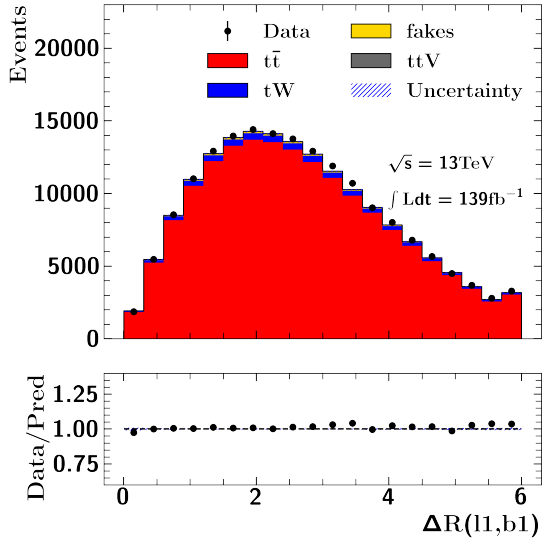
A Appendix

A.1 Kinematic properties (Section 5.4)



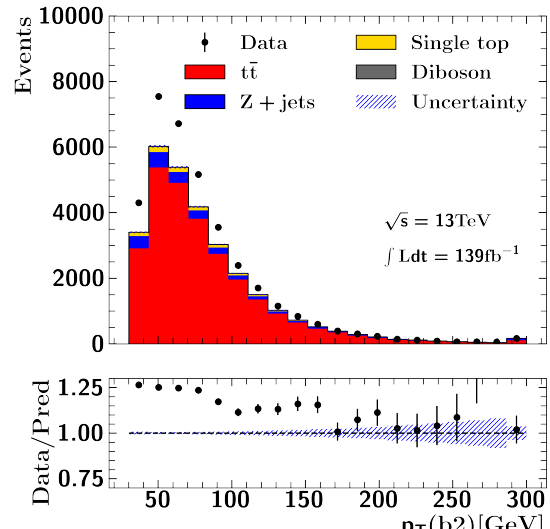
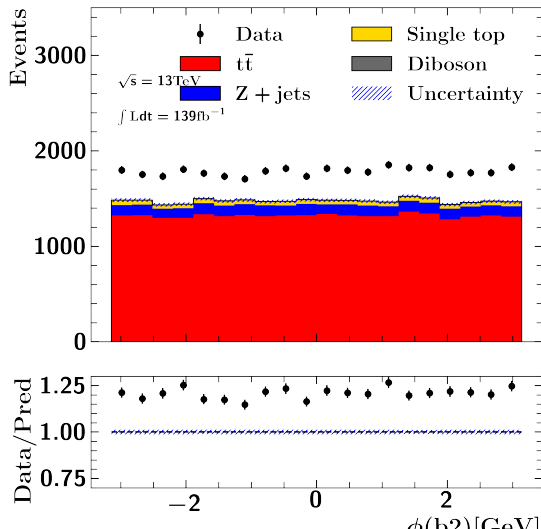
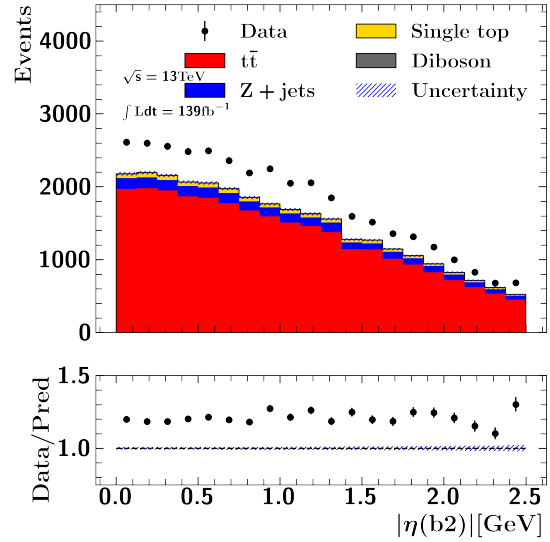
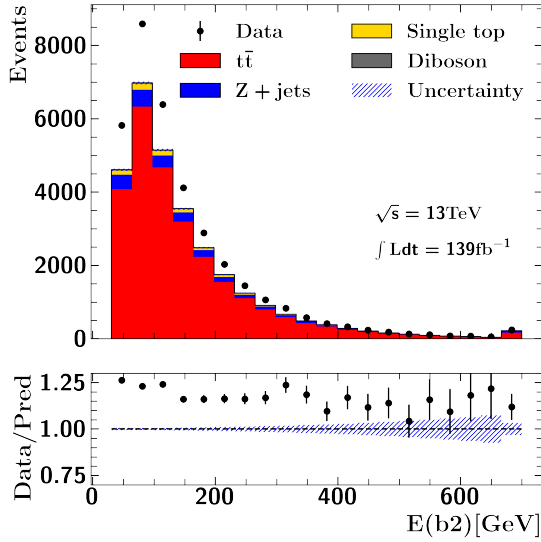
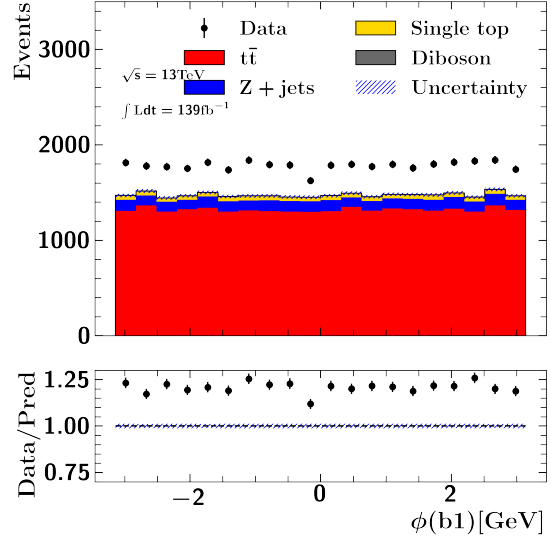
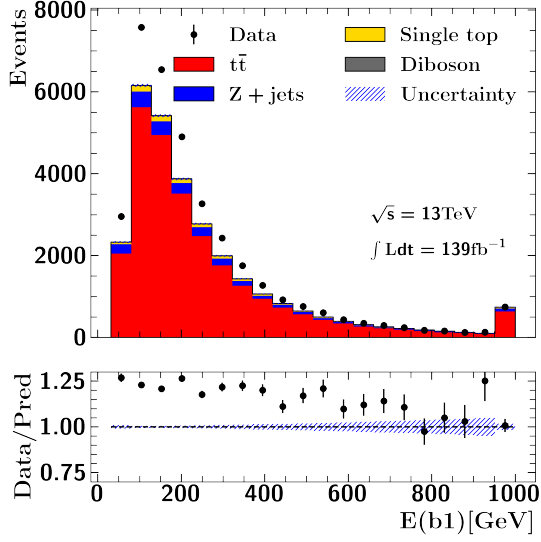


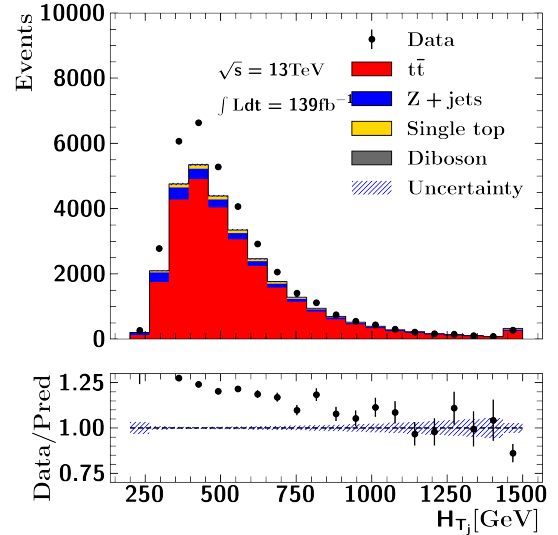
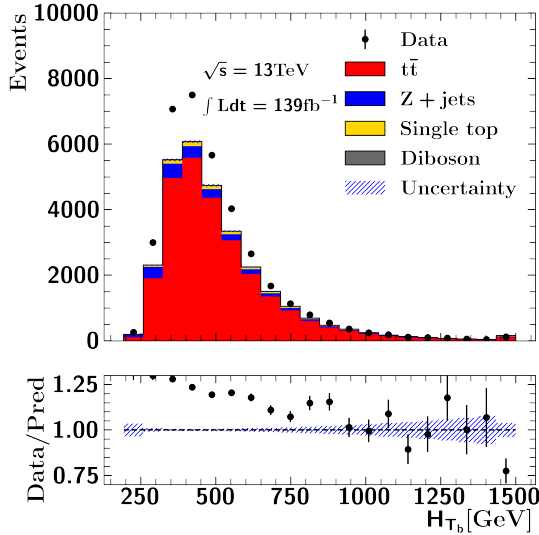
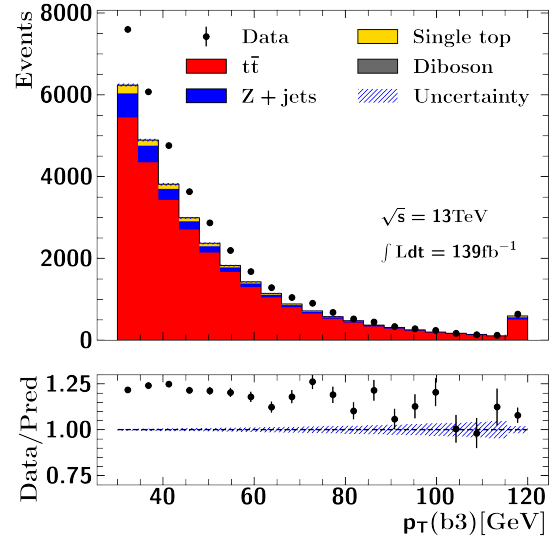
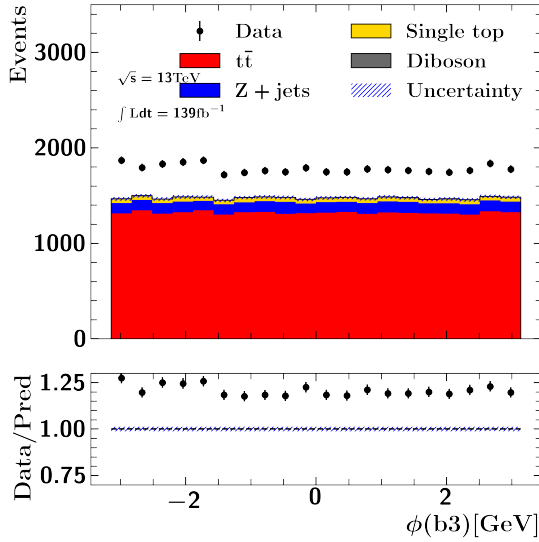
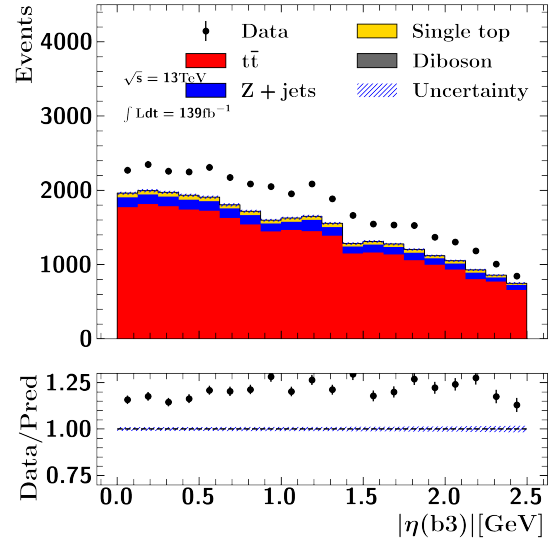
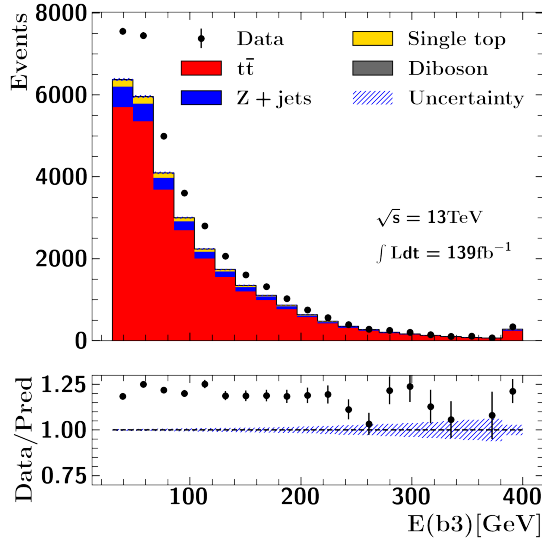


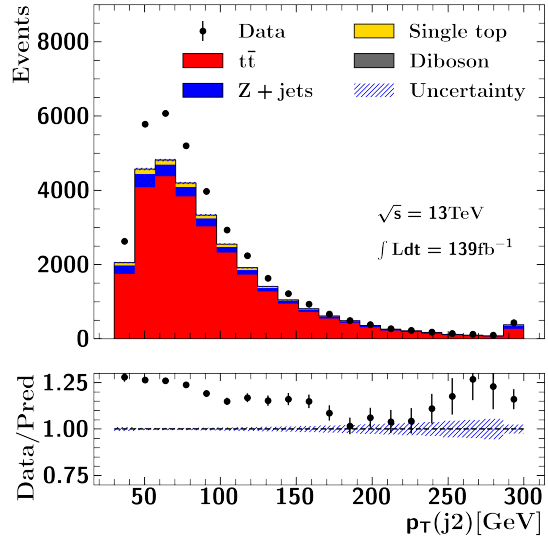
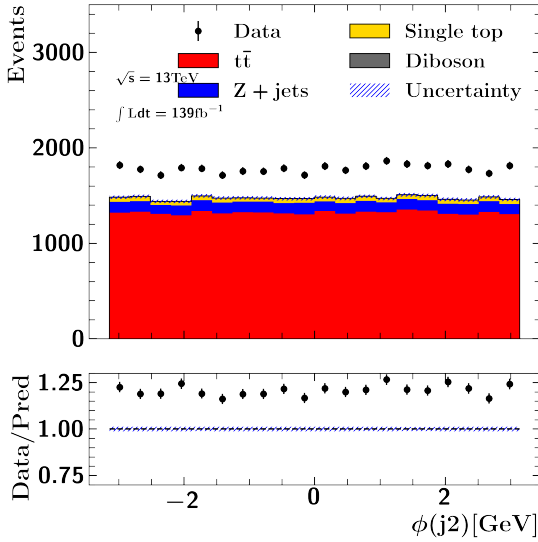
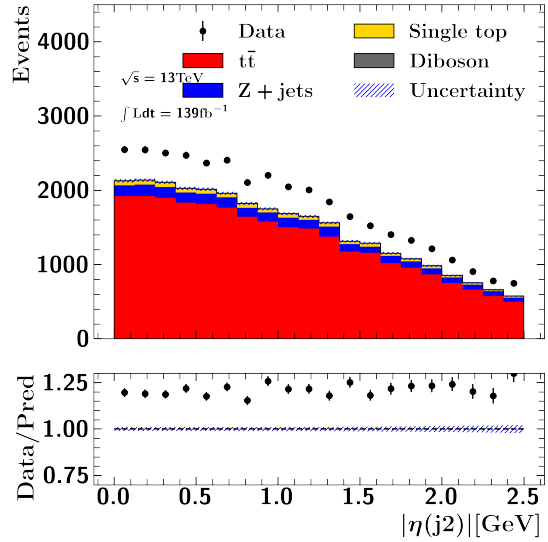
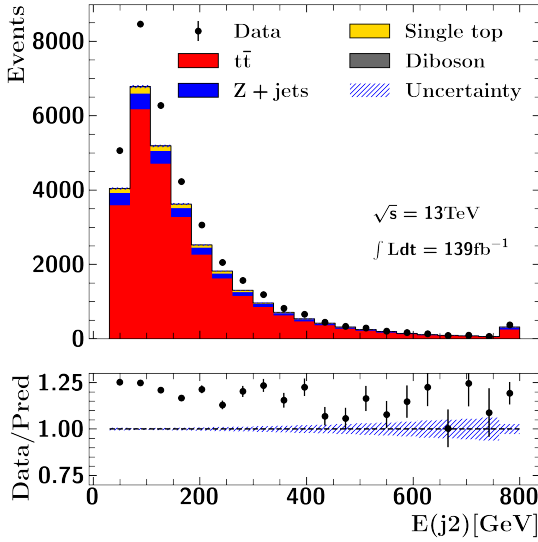
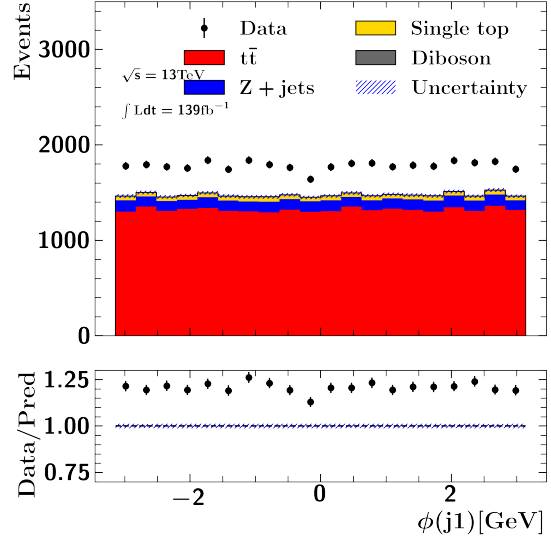
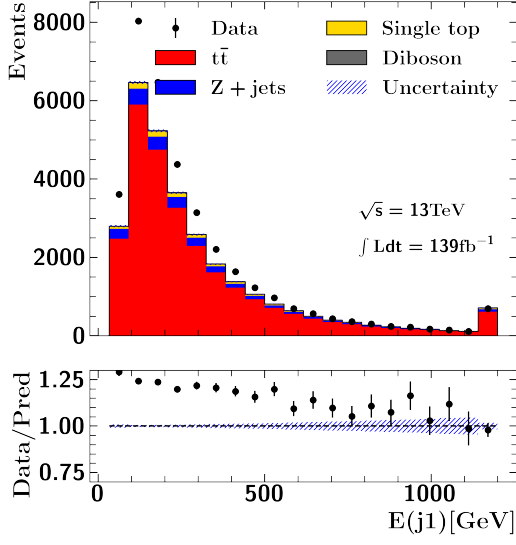


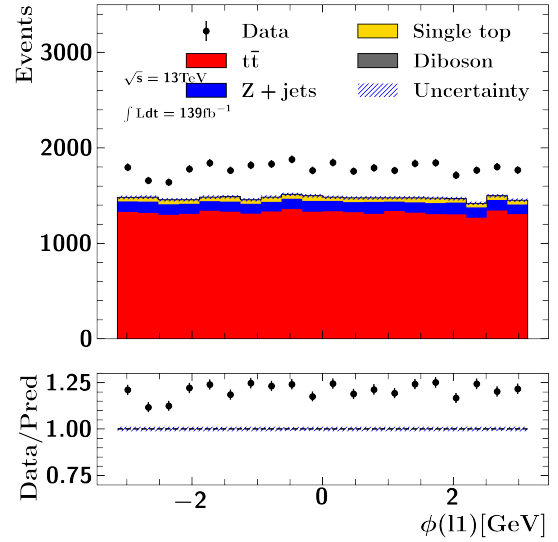
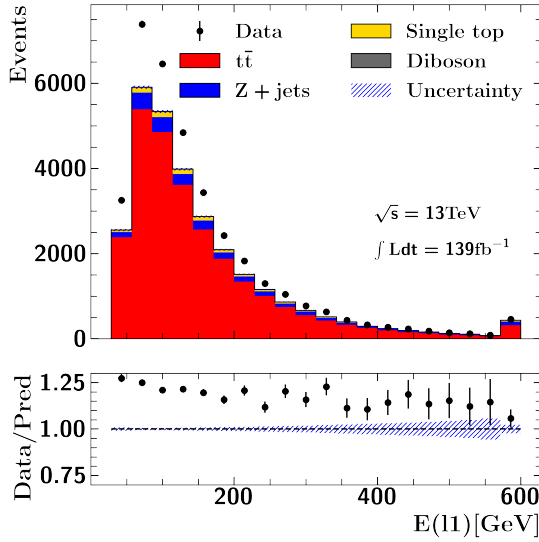
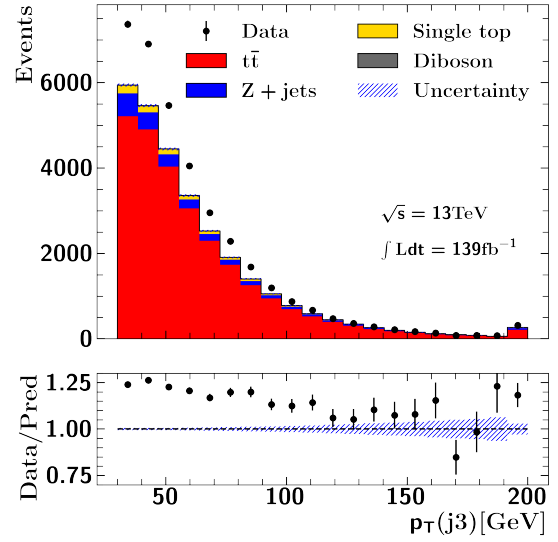
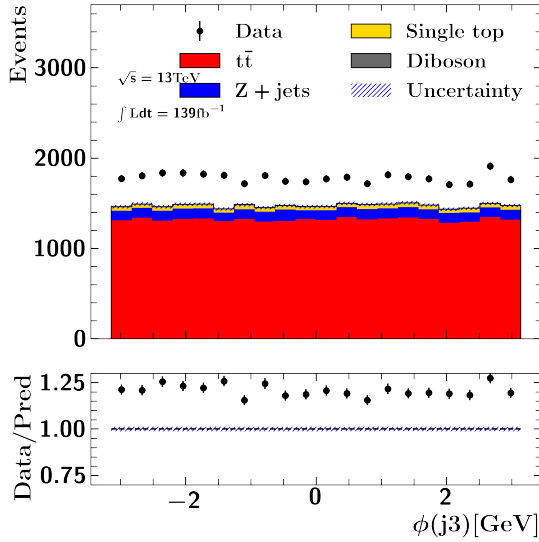
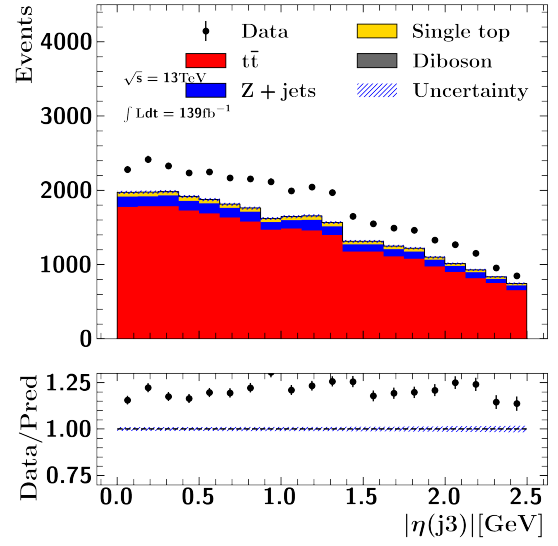
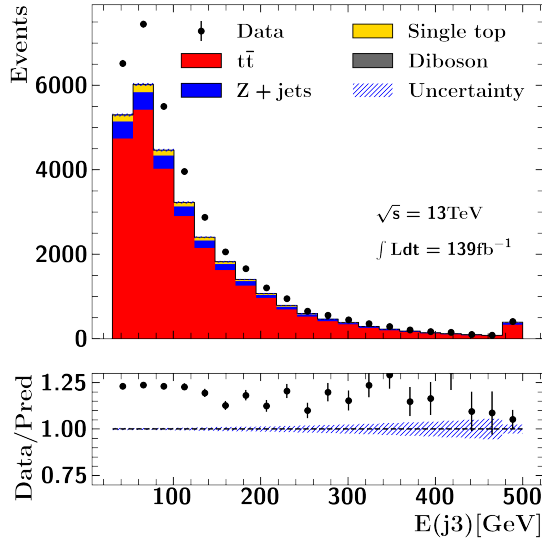
B Appendix

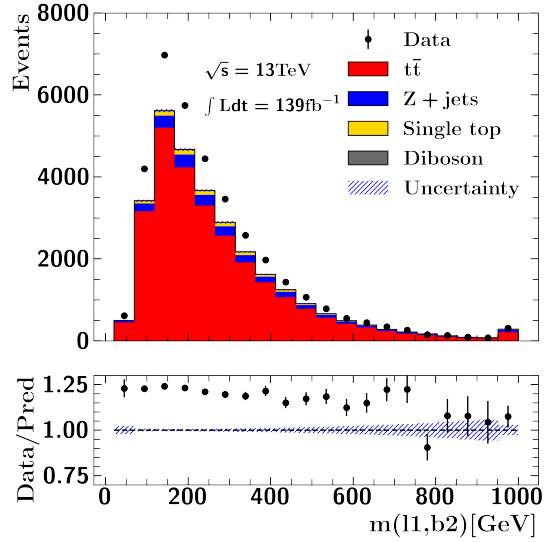
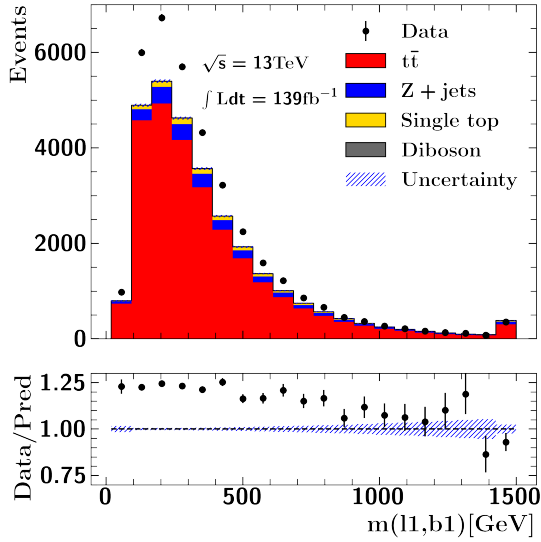
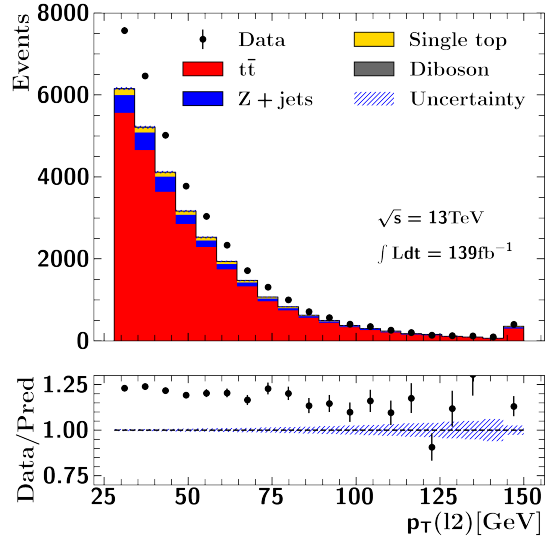
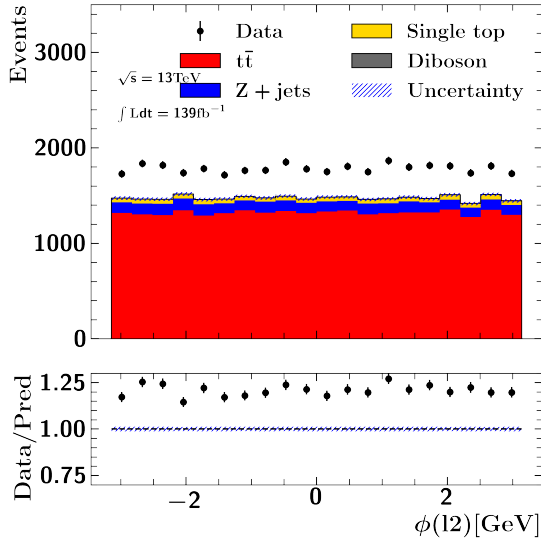
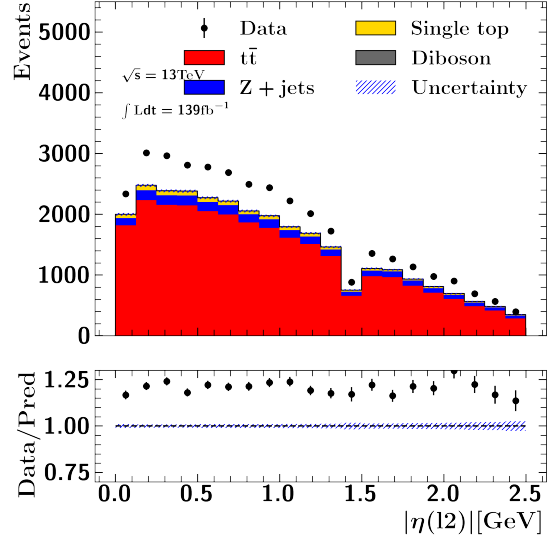
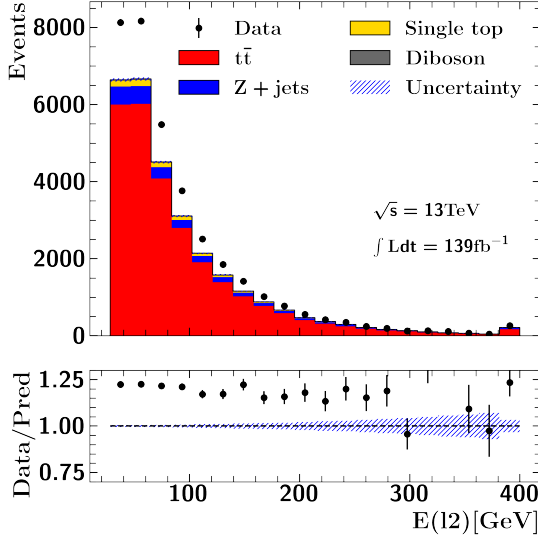
B.1 Kinematic properties (Section 6.3)

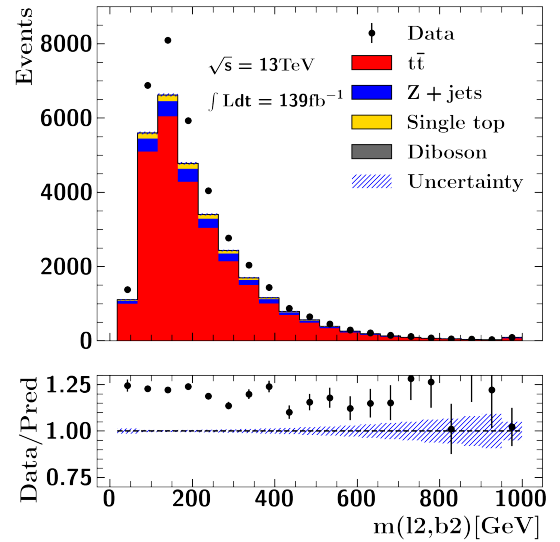
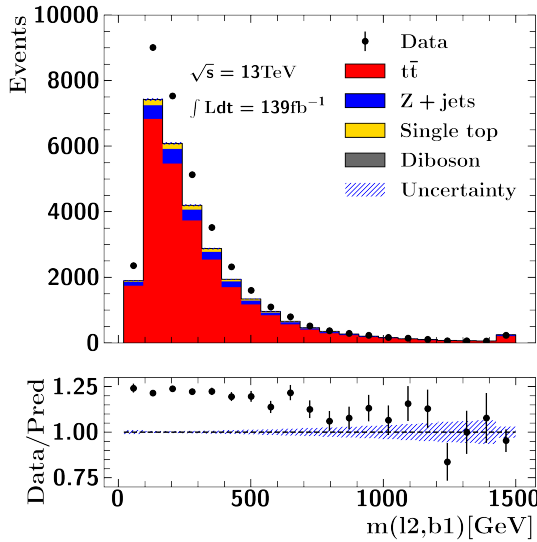
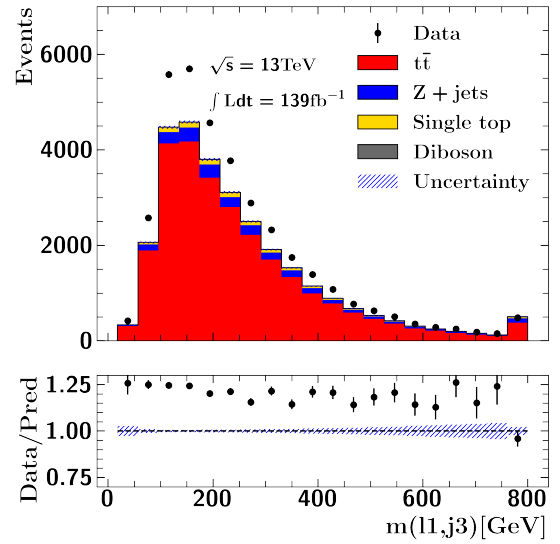
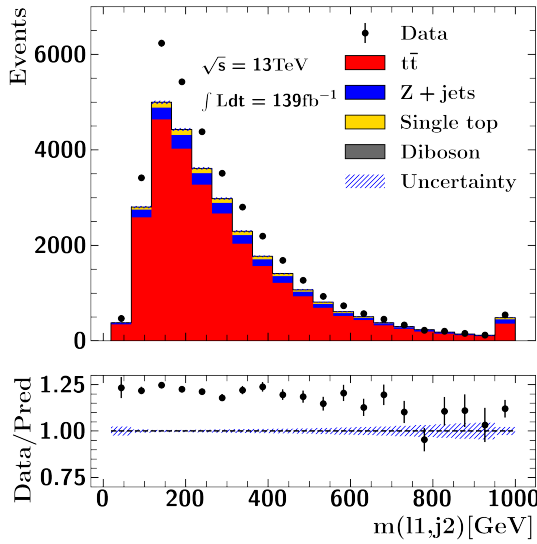
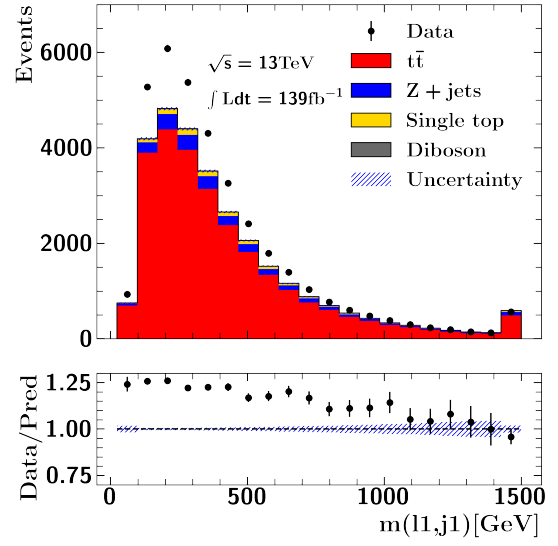
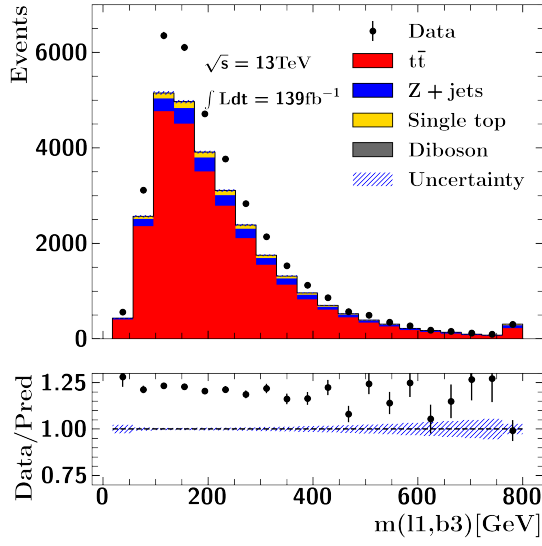


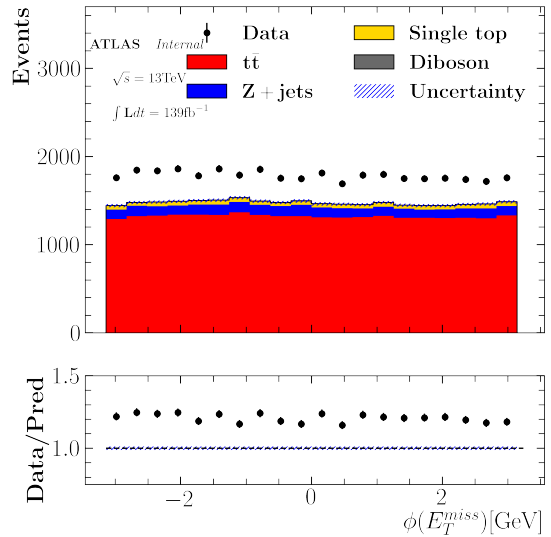
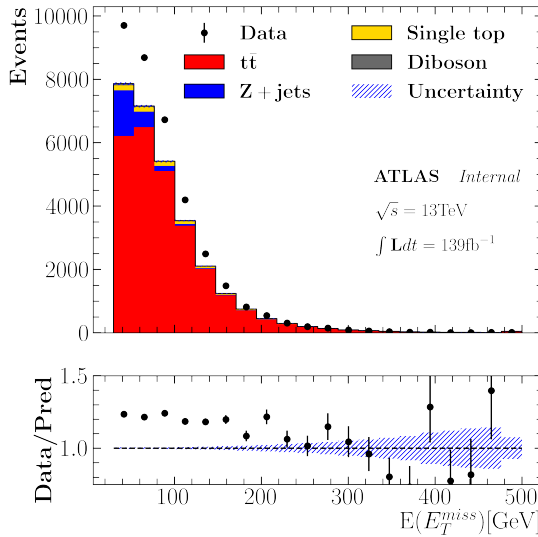
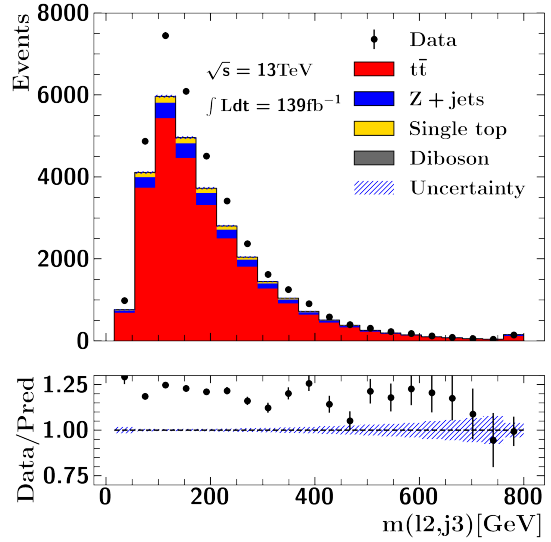
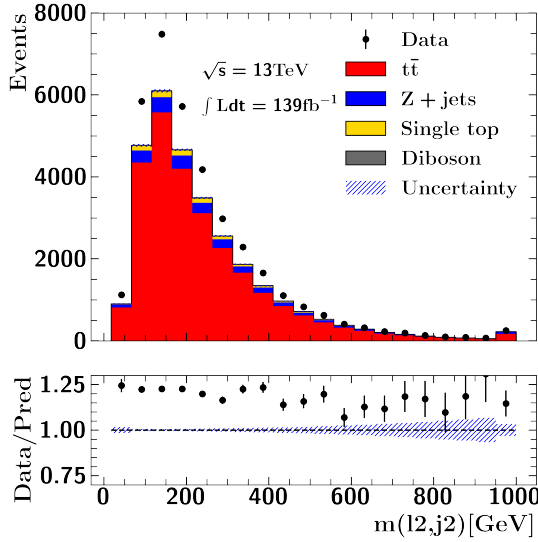
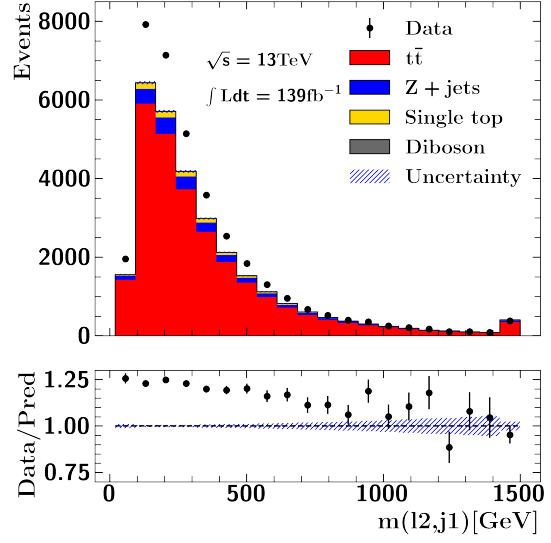
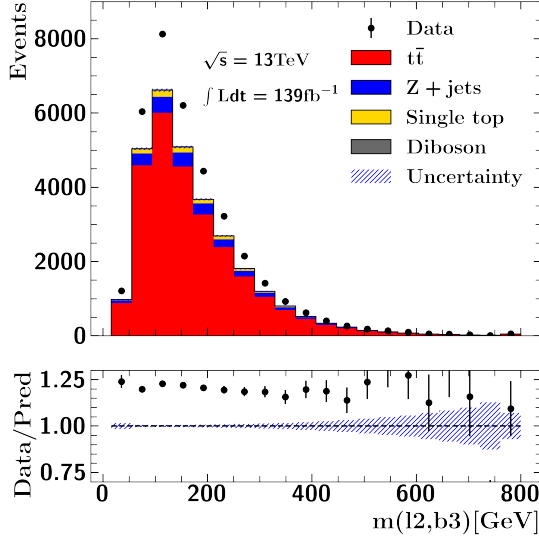


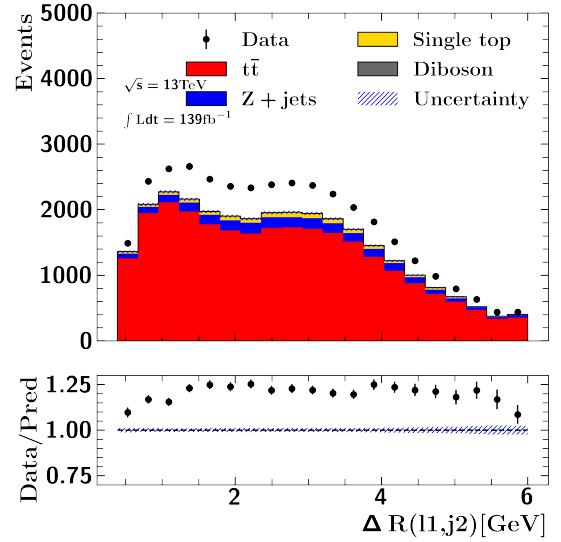
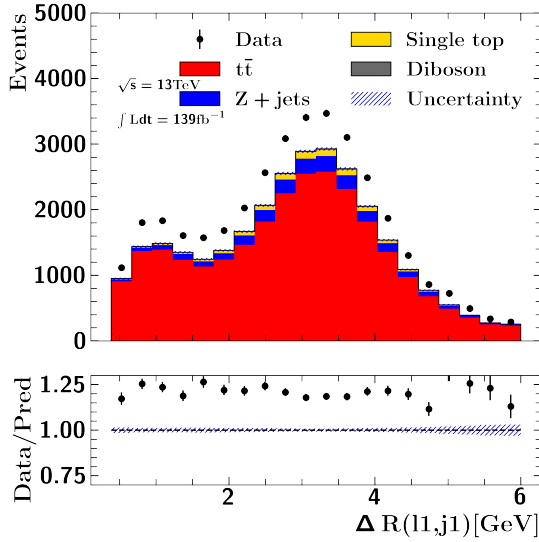
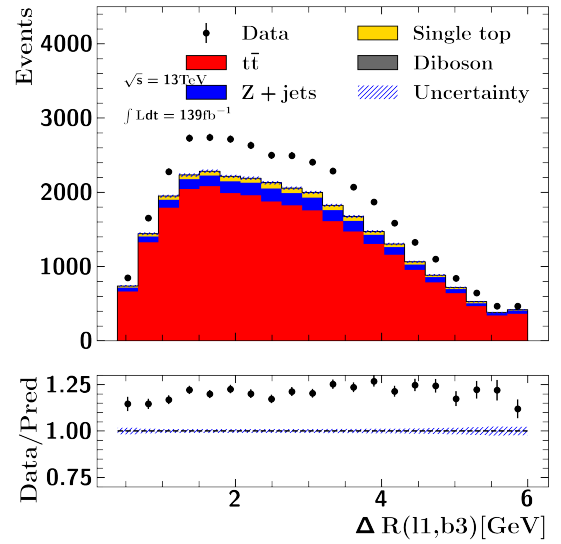
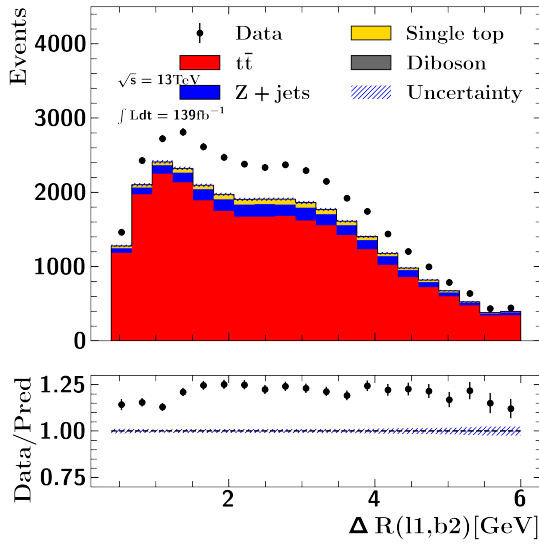
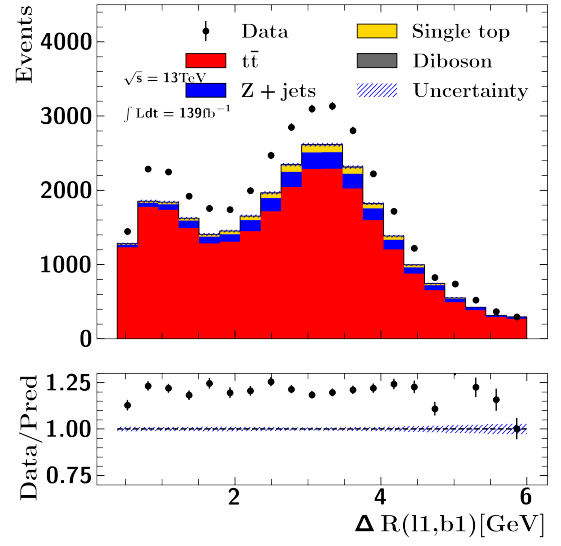
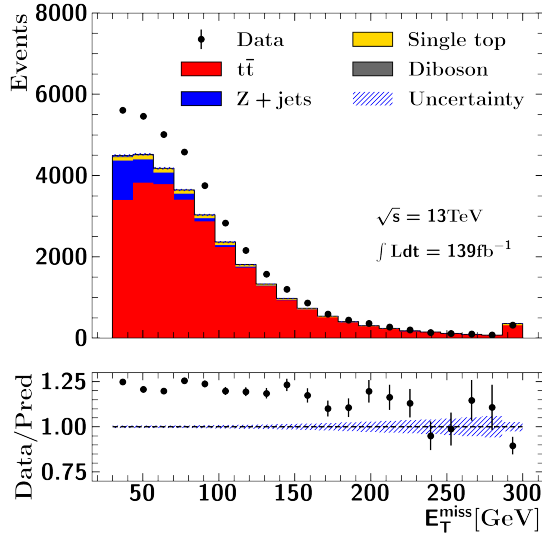


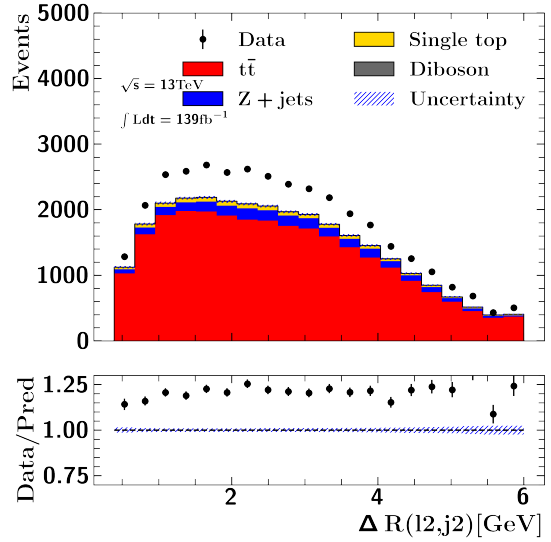
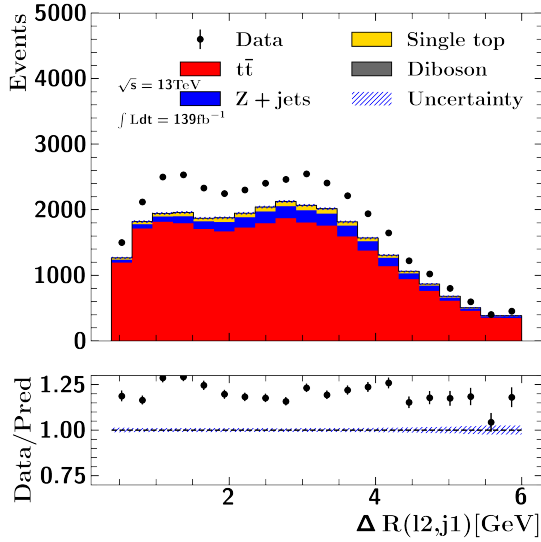
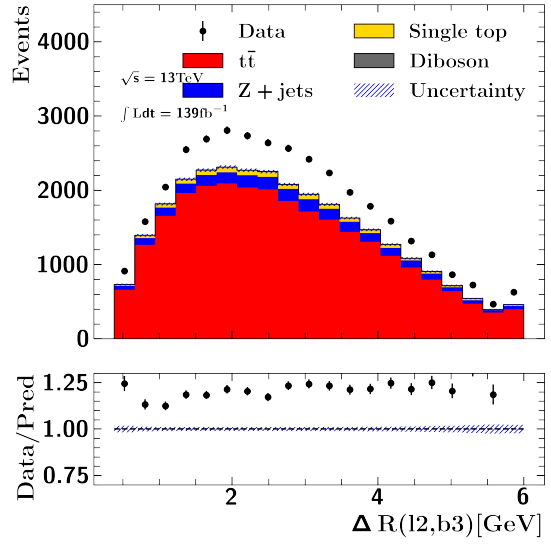
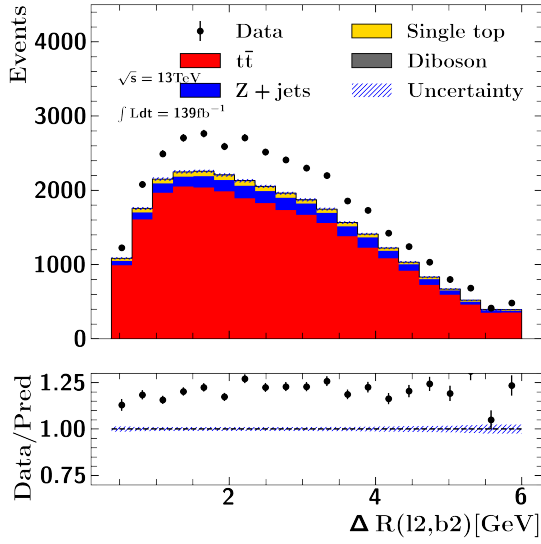
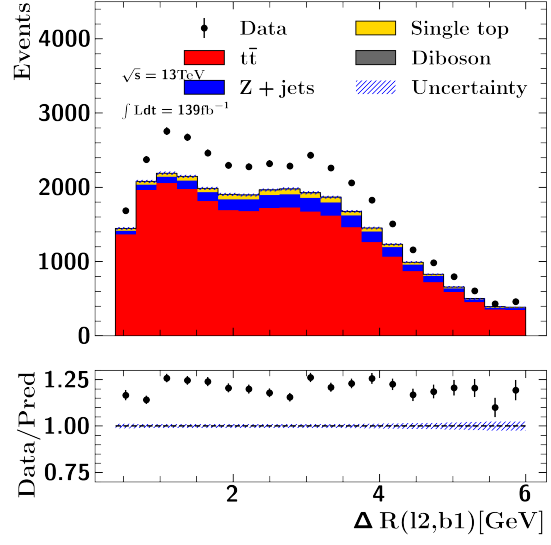
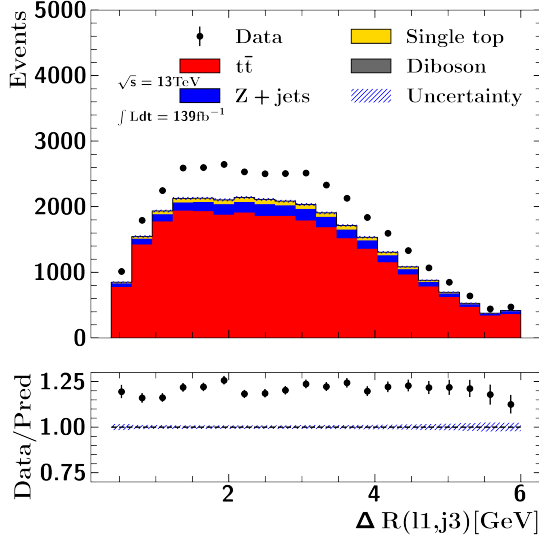


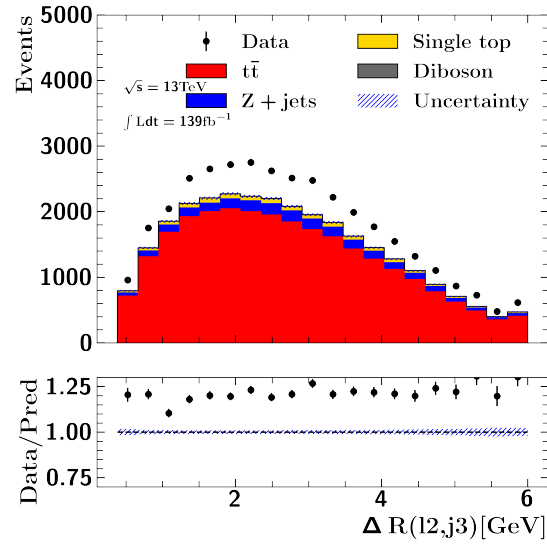












Name, Vorname:

Anschrift:

.....

E r k l ä r u n g

gem. § 20 Abs. 9 PO

Hiermit erkläre ich, dass ich die von mir eingereichte Abschlussarbeit (Master-Thesis) selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die Stellen der Abschlussarbeit, die anderen Werken dem Wortlaut oder Sinn nach entnommen wurden, in jedem Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe. Entsprechendes gilt für beigegebene Zeichnungen, Kartenskizzen und Darstellungen.

Sollten entsprechend der Themenstellung ggf. Vorarbeiten des Forschungsprojektes in die Abschlussarbeit eingeflossen sein, so habe ich dieses gekennzeichnet bzw. als Anhang nachgewiesen.

.....
Datum

.....
Unterschrift

E r k l ä r u n g

Hiermit erkläre ich mich damit einverstanden, dass meine Abschlussarbeit (Master-Thesis) wissenschaftlich interessierten Personen oder Institutionen zur Einsichtnahme zur Verfügung gestellt werden kann.

Korrektur- oder Bewertungshinweise in meiner Arbeit dürfen nicht zitiert werden.

.....
Datum

.....
Unterschrift